STREPTOMYCES EVOLUTION AND BACTERIAL SPECIATION THEORY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

James Robert Doroghazi

January 2011

STREPTOMYCES EVOLUTION AND BACTERIAL SPECIATION THEORY

James Robert Doroghazi, Ph.D.

Cornell University 2011

Classification of bacterial species and studying bacterial speciation are both difficult and necessary. *Streptomyces* is a genus of Actinobacteria; they have 7-10Mbase high G+C genomes, and produce the majority of clinically useful antibiotics. Genome sequencing has revealed overlooked biosynthetic potential, making many hopeful for a genome based surge in natural product discovery. Unfortunately, *Streptomyces* diversity is not well understood, and while there are 376 recognized species, many studies of the genus discuss 20 or so major clusters. Multilocus sequence analysis (MLSA) has been used recently to establish a foundation for the study of species and diversity within *Streptomyces*. We found significant interspecies recombination within one 53 species MLSA data set that affects >40% of the concatenated six gene sequences. An MLSA data set was also created for 38 isolates of *Streptomyces flavogriseus* phylogroup *pratensis*. The homologous recombination (HR) rate for this population was found to be very high for bacteria. We used the same 38 isolates to determine the uniformity of phenotypic traits within one *Streptomyces* population. We found that while carbon-source utilization was low compared to other bacterial species, it is not entirely uniform and is therefore not a diagnostic trait for *Streptomyces* species. Secondary metabolite production gene clusters were, however, very well conserved within this population. In a broader study of *Streptomyces* diversity, we examined streptomycetes isolated from 15 soil samples spanning Florida to Alaska. We found that it is possible to easily delineate *Streptomyces* species using

MLSA data for populations. The species studied had very different recombination rates, ranging from clonal to highly recombinant. The theoretical study included in this dissertation addresses the effect of homologous recombination rate on divergence of bacterial populations. We have found that homologous recombination acts as a cohesive force, holding two populations together that would otherwise be on separate evolutionary trajectories. This cohesion is rate dependent, and because HR rate decreases as sequence divergence increases, there is a threshold above which there is no net population cohesion. This may result in two different sets of considerations when studying microbial ecology and diversity.

BIOGRAPHICAL SKETCH

James Robert Doroghazi was born on July 8, 1984 to Paul and Mickey Doroghazi in Sterling, Illinois. His father is a general surgeon, and his mother is a nurse. He has two sisters, one older and one younger. James graduated first in his class from Garber High School in Essexville, Michigan in 2002. James was originally drawn to research after his freshman year at Michigan State University. Prior to this he had been considering veterinary medicine, but after spending several days shadowing a veterinarian, the prospect of performing the same tasks day after day for an entire career made him reconsider. To James, research was the most obvious choice of a career that would constantly present new and challenging ideas. James began work in the summer of 2003 in the lab of Dr. Steve Triezenberg in the virology department of MSU, where he worked until his graduation three years later. In Dr. Triezenberg's lab, James studied gene expression of herpes simplex virus type 1. After graduation from MSU, he married Heidi Schanhals, his girlfriend of three years. Together they moved to Ithaca, NY so that James could attend graduate school. At Cornell, James became fascinated with the study of bacterial biogeography. *Streptomyces* as the focus of a biogeography study was suggested by Dr. Dan Buckley, his graduate advisor. The focus of this work shifted as it was realized that prior to studying the biogeography of streptomycetes, a species framework as a foundation for the study of their diversity would be required. While in graduate school, James and his wife have had two daughters, Hannah and Olivia. He plans to pursue postdoctoral research in the Mining Microbial Genomes for Novel Antibiotics theme at the University of Illinois Institute for Genomic Biology.

To my wife and daughters, for making life beautiful.

ACKNOWLEDGMENTS

I would like to first thank my advisor, Dan Buckley, for giving me the freedom to pursue this project and for the excellent suggestion to study *Streptomyces*. Ashley Campbell and Pete Kelly have contributed significant work to this dissertation: Ashley to Chapter 5, and Pete to Chapters 4 and 5. My two minor advisors, Steve Zinder and Carlos Bustamante, have provided valuable insight and observations. Rose Loria and two members of her lab in particular, Dawn Bignell and Jose Carlos Huguet, have given me very valuable advice on *Streptomyces* and suggestions for my research. Peter Bergholz has provided useful comments and suggestions on this research, especially on the biogeography work presented in Chapter 5. I would also like to thank everyone who has provided us with soil samples, a list that is too long to include here. This work has been supported in part by the Cornell Center for Comparative and Population Genomics in the form of a graduate student fellowship.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Bacterial Species Concepts

A unified microbial species concept is one of the most divisive topics in all of microbiology. One of the reasons for this is an inability to squarely define a fundamental unit of bacterial diversity, be it a species, ecotype, or population. The field of bacterial systematics has been, and remains to be, about finding clusters of bacteria that are more closely related to each other than to outlying groups. Phenotypic clustering was the only method available to early systematists (e.g. (1)). Recreating these phenotypic clusters using molecular data, such as DNA-DNA hybridization, was the guide for the creation of currently used species delineation criteria (2-4). 97% 16S rRNA gene similarity was later found to correspond roughly to this DNA-DNA hybridization value, although this 16S and DNA-DNA hybridization relationship is intended to be interpreted in one direction: if <97% 16S rRNA, then almost surely <70% DNA-DNA hybridization, as high rRNA similarity is possible in organisms with low overall DNA similarity (5).

Studies on mechanisms of bacterial speciation have focused on two major forces, homologous recombination and ecotype formation, and emphasis on one over the other creates the main viewpoints currently held by most about bacterial species concepts. The ecotype concept places an emphasis on periodic selection as a means to prevent diversification within an ecologically homogeneous group. New ecotypes are formed when a new niche is invaded by the ecotype founder (6). Homologous recombination (HR), here defined as the transfer of homologous DNA between a donor and a recipient, is commonly referred to as a “cohesive force” and is the second

focus of bacterial speciation studies. HR has been shown in multiple models to prevent cluster formation within a species when the recombination rate is above a somewhat nebulous threshold (7, 8). Chapter 2 of this dissertation quantifies this cutoff and gives equations for convergence or divergence rates between separate bacterial populations. This theoretical advance has important implications, uniting work that otherwise appears contradictory.

Population Genetics of Bacteria

Bacteria reproduce through binary fission, with mutation as the only source of change in the vertical inheritance of genetic material from the mother cell to the daughter cells. HR is generally as likely to affect a site as mutation, a finding replicated in multiple systems (9). Given the similarity between the likelihood of recombination or mutation to affect any given site (r/m), recombination is now rightly considered an important force to consider when examining bacterial populations.

Bacterial recombination can be divided into two classes: homologous (HR) and nonhomologous (NR). Non-homologous recombination involves the lateral transfer of DNA through a variety of mechanisms from a donor to a recipient that, prior to the event, did not have a homologous region of DNA. This is probably the more commonly discussed type of lateral gene transfer (LGT) because it spreads accessory genes, allowing new functions to be performed by the recipient. This can result in niche expansion or the transfer of "weapons" in a coevolution scenario, examples of which are transfers of pathogenicity islands and antibiotic resistance. Nonhomologous recombination can therefore be seen as a primarily diversifying force, allowing populations to branch off and expand in new directions (10) and can have a profound

impact on genomic diversity. In 61 sequenced *Escherichia coli* genomes, 80% of the genes in any one isolate will not be found in all of the remaining 60 (11). Genes present in all strains make up the core genome, which asymptotically shrinks as more genomes are examined . This is contrasted with the pan-genome, which must be finite as there are not infinitely many individuals in any species, but may be practically infinite in terms of our ability to sample new genes (12). One problem with most of the studies on this topic to date is that genome sequencing is not yet cheap enough to encourage random sampling of strains to study. When choosing strains that are representatives of different groups, one should expect to find novel genetic diversity in each new genome.

Homologous recombination is the one-way transfer and incorporation of similar DNA into the recipient's chromosome, and is in this way more of a gene conversion event than a recombination event in comparison with eukaryotic genetics. Homologous recombination rates vary based on the level of homology between the donor and recipient DNA (13-16). The relationship between divergence and recombination rate is best fit with a log-linear curve, described by the equation: $c = R*10^{-20d}$, where c is the real per site recombination rate, R is the baseline recombination rate when sequences are identical, d is the divergence between sequences, and 20 is the distance factor, such that when sequences are identical d = 0 and c = R (7). This is estimated at the population level for haploid species as $\rho = 2N_e c$, where N_e is the effective population size. The per site mutation rate is μ, which is calculated at the population level as $\theta = 2N_e\mu$. A commonly reported value is ρ/θ; this value differs from r/m, as ρ/θ is a ratio of rates and r/m is the ratio of sites affected by recombination versus sites affected by mutation. The value for r/m is commonly higher than ρ/θ, as each recombination event

is more likely to introduce more changes than each mutation event.

Models of Bacterial Populations

To create expectations of the patterns and processes that recombination and mutation produce at the population level, several models of bacterial populations have recently been created. Fraser *et al*. have created a neutral microepidemic model, which solves for the equilibrium expression of the allelic mismatch distribution, which is similar to heterozygosity in diploid population genetics models. The equilibrium state of this model fits the data from surveys of *Streptococcus pneumoniae*, *Neisseria meningitidis*, and *Staphylococcus aureus* isolates well (17). The same group has also created an infinite allele models modeling a bacterial population with a set of 70 loci, including in this model a reduction in recombination rate due to increasing genetic dissimilarity. While cluster formation was observed at high recombination rates and with unrealistically high distance factors, permanent cluster formation was not observed with parameters close to those estimated from recombination experiments (7, 18).

Vetsigian and Goldenfeld created a similar model which included a small population of circular genomes allowing for mutation and homologous recombination, with and without the presence of an initial nonhomologous region, similar to what would be seen after an illegitimate recombination event. They observed that a nucleus of differences can serve as a starting point for "propagating fronts" of divergence, depending on mutation and recombination rates, as well as the recombination distance factor (8). Falush *et al*. implemented a model similar to those just described but differing in the parameter values used, most importantly in the recombination distance factor. Their simulation resulted in speciation within 500-1,000 generations in a small

population of 500 to 1,000 individuals (19). The recombination distance factor used, however, was the same unrealistic factor used by Fraser *et al*. to achieve cluster formation; this is close to 15 times the distance factor observed in laboratory experiments.

Field Studies of Bacterial Populations

Multilocus sequence typing (MLST) is currently the tool of choice to measure real world bacterial population variation and structure (20). The basis of this approach is to sequence a set of genes spread across the chromosome. As it was originally developed for typing of pathogens, MLST is useful for identifying different strains of closely related bacteria. This method provides more information and is more sensitive than its predecessor, multilocus enzyme electrophoresis, which relied on differences in protein structure and amino acid composition to define alleles of the chosen enzymes. MLST instead uses DNA sequence; this provides more information and allows the investigator to test for neutrality in the chosen genes using ratios of synonymous and nonsynonymous polymorphisms. Various MLST schemes have been designed for different groups of bacteria, with some species being investigated using multiple schemes; while the design flaws and achievements of an individual scheme can be debated, it is more useful overall for all researchers working within a group to use the same scheme. This allows direct comparisons of different studies, and creation of large datasets added to by multiple groups (20, 21).

Most of the population level studies of bacteria have been performed on pathogens. Two of the most studied examples are *Neisseria meningitidis* and *Helicobacter pylori*. Both of these bacteria have high recombination rates in part because they are naturally

transformable; they are capable of taking up DNA present in the environment, for nutrition or recombination. The program LDHAT was used to analyze population data for *Neisseria meningitidis* and estimated that ρ/θ is ~1 (22). Interestingly, the same survey found that the alleles appear to be shared in a common gene pool, but that the combination of the alleles were different between disease and carriage isolates. H. pylori has been found to recombine so frequently that over 41 years the equivalent of half of the genome may be imported and replaced by recombination (23).

Two studies have reanalyzed data from many MLST surveys to allow for uniform comparisons across populations. The first of these, by Perez-Losada *et al*., found *Neisseria meningitidis*, *Neisseria gonorrhoeae*, *Helicobacter pylori*, and *Streptococcus pneumoniae* had the highest levels of intragenic recombination among the species examined using the program LDHAT for analysis (24). Vos and Didelot found the same species to have high r/m values using the program CLONALFRAME (25).

Of the studies not performed on pathogens or host-associated bacteria, many have been done on hot spring or saltern pond populations, as these have discrete islands of identifiable habitat and smaller population sizes. Papke *et al*. found clustering within hot spring cyanobacteria based on geographic location, with no correspondence to chemical characteristics within the sites surveyed (26). Isolation by distance was found in hyperthermophilic archaea, also inhabitants of hot springs (27). Within this same system, recombination has been found to have a strong effect on population structure resulting in random association between alleles at different loci, allowing for genetic diversity to persist despite selective sweeps (28). Haloalkaliphilic bacteria in

the genus *Thioalkalivibrio* isolated from soda lakes in Asia, Africa, and North America show clustering based on broad geographic regions, but most genotypes, as characterized by repetitive extragenic palindromic PCR (REP-PCR, a genomic "fingerprinting" technique), were found in only one area (29). *Halorubrum* populations exhibit linkage equilibrium almost to the extent of sexual populations (30). Several haloarchaea even contain multiple highly divergent copies of 16S rRNA (31). Some consider these systems ideal model systems to study speciation in Archaea, due to high expression of recombination systems, high diversity, islands of habitat, and dynamic genomes (26). These same traits also limit the applicability of the studies and raise questions about how functions studied in standard lab systems may differ to a significant degree in these extreme environments.

There has also been a small number of studies performed on *Streptomyces*, although their primary goal was not necessarily to assess the biogeography and structure of one population. Davelos *et al*. attempted to examine composition of *Streptomyces* communities in prairie soil, using 16S rDNA sequences to cluster isolates collected from three randomly chosen locations within a 1 m^2 area. 26 of the 34 operational taxonomic units (OTU) recovered were singletons or doubletons, with the most abundant OTU represented by 55 out of 153 isolates. They reported significant differences between the three locations and at different depths (32). Differences between sites may have been caused by the marker of choice not providing significant resolution and perhaps by the shallow depth of their sampling. Antony-Babu and Goodfellow found that 16S rRNA genes from *Streptomyces* isolated from a beach and dune sand system formed clusters that can be equated with species. They also found a perfect correlation between colony color and REP-PCR groups, in that every member

of any given REP-PCR group belonged to the same colony color group (33).

A broader investigation of actinomycete communities, including *Streptomyces*, found that 16S rRNA gene fingerprints and polyketide synthase genes clustered by region of origin, indicating different or genetically distinct actinomycetes present in North American and Asian soils (34). While this makes the hypothesis that *Streptomyces* subpopulations may be isolated by distance plausible, a cosmopolitan distribution cannot yet be ruled out. Hints of regional endemicity have also been seen in plant pathogenic scab causing *Streptomyces* (35-38). Isolation by distance has also been recorded for another spore-forming bacterium, *Myxococcus xanthus* (39).

Streptomyces Biology

Streptomyces do not fit the general bacterial mold in terms of biology, recombination or genomes. Originally classified as a type of fungus (40), *Streptomyces* share several morphological and ecological traits with their eukaryotic doppelgangers. *Streptomyces* growth strategies are ideal for penetrating through solid substrate, most commonly plant matter. Growth on a surface starts with germination of a spore into hyphae, threadlike growths capable of branching, which make up the substrate mycelium. The next stage of growth is the formation of aerial hyphae, growing up off of the substrate and therefore relying on the substrate mycelium for nutrients. Aerial hyphae lead to, quite literally, the seminal accomplishment of the *Streptomyces* colony, the production of more spores. Spore formation occurs at the ends of the aerial hyphae; when hyphal growth is complete, the tips septate into compartments that mature into spores (41). These spores are not true endospores, and as such are not exceedingly heat resistant. The purpose of the spores appears to be dessication resistance, allowing prolonged

dormancy and ideal conditions for wind dispersal. Outside of the spore stage *Streptomyces* hyphae and mycelia are dissimilar from most bacterial cells in that they are not unigenomic compartments; instead of a single genome physically separated from others by cell walls, around ten *Streptomyces* genomes share a compartment, separated from the next compartment by a septation (40, 41).

Recombination also occurs quite differently in *Streptomyces*. For the transfer of genetic material, hyphae must fuse. This is thought to occur only at the hyphal tips, as that is the location where new growth occurs. It is possible that during the transfer of DNA whole genomes are transferred, although the mode of transfer and the extent of DNA transferred are not yet sufficiently resolved. Unlike F-mediated recombination, double stranded DNA is transferred to the recipient, as expression of a double strand restriction enzyme system in the recipient destroys the ability to recombine. Recombination is thought to be mediated by plasmids, although other possibilities cannot be ruled out (40, 42).

The genomes of *Streptomyces* also differ from most bacteria in several ways. *Streptomyces* genomes are large, 7-9Mb or more, and linear, with replication proceeding bidirectionally from a centrally located origin of replication[33]. The core genes tend to be located in the middle of the linear molecule, with accessory genes closer to the ends. Within the core genes, synteny has been remarkably well preserved among the *Streptomyces* sequenced to date (43). Outside of the conserved core, amplified segments and deletions occur more frequently. This is possibly to balance the chromosome arms so that the ori remains in the middle (44).

Streptomyces can degrade many various substrates but are also capable of producing a large variety of secondary metabolites. Most of the clinically important antibiotics are produced by *Streptomyces* (45). The diversity of secondary metabolites is partly due to the number of species and amount of diversity contained within the genus, but each individual is also capable of creating a variety of products; in the *S. coelicolor* genome sequence there are 23 biosynthetic gene clusters involved in secondary metabolite production (45). This should probably be viewed as a lower estimate, as it only counts genes known to be typical of secondary metabolism. It may be that the ability to produce 10-20 secondary metabolites is average among actinomycete strains (46). Not all of these metabolites are produced under any one set of conditions, allowing genome sequencing to reveal metabolites previously unidentified due to incorrect expression conditions in regular lab media. Enzyme structure predictions based on DNA sequence has allowed for more accurate product structure predictions, exemplified by a novel antifungal agent that was correctly predicted from the genome sequence of *Streptomyces aizunensis* (47). Systematic approaches will be crucial to discovery of novel metabolites, as searching without a plan is never the most efficient method. Donadio *et al*., describing work at Biosearch Italia state: "Isolation methods are usually applied to soil samples or other specimens in a random way…. It would be highly desirable to know in advance…their relative abundance and, possibly, their diversity, so that an appropriate effort can be devoted to that source". (48)

REFERENCES

1. Jones D, Sackin M, & Sneath P (1972) A numerical taxonomic study of streptococci of serological group D. *Microbiology* 72(3):439.

2. Palleroni N, Kunisawa R, Contopoulou R, & Doudoroff M (1973) Nucleic acid homologies in the genus *Pseudomonas*. *Int. J. Syst. Evol. Microbiol.* 23(4):333.

3. Johnson J (1973) Use of nucleic-acid homologies in the taxonomy of anaerobic bacteria. *Int. J. Syst. Evol. Microbiol.* 23(4):308.

4. Cohan FM & Perry EB (2007) A systematics for discovering the fundamental units of bacterial diversity. *Curr. Biol.* 17(10):R373-386.

5. Stackebrandt E & Goebel B (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* 44(4):846.

6. Cohan FM (2002) What are bacterial species? *Annu. Rev. Microbiol.* 56:457-487.

7. Fraser C, Hanage WP, & Spratt BG (2007) Recombination and the nature of bacterial speciation. *Science* 315(5811):476-480.

8. Vetsigian K & Goldenfeld N (2005) Global divergence of microbial genome sequences mediated by propagating fronts. *Proc. Natl. Acad. Sci. U. S. A.* 102(20):7332-7337.

9. Vos M & Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 3(2):199-208.

10. Lawrence JG (2002) Gene transfer in bacteria: speciation without species?

Theor. Popul. Biol. 61(4):449-460.

11. Lukjancenko O, Wassenaar TM, & Ussery DW (2010) Comparison of 61 Sequenced *Escherichia coli* Genomes. *Microb. Ecol.* Epublished ahead of print.

12. Tettelin H*, et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.* 102(39):13950-13955.

13. Shen P & Huang HV (1986) Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics* 112(3):441-457.

14. Zawadzki P, Roberts MS, & Cohan FM (1995) The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust. *Genetics* 140(3):917-932.

15. Zahrt TC & Maloy S (1997) Barriers to recombination between closely related bacteria: MutS and RecBCD inhibit recombination between *Salmonella typhimurium* and *Salmonella typhi*. *Proc. Natl. Acad. Sci. U. S. A.* 94(18):9786-9791.

16. Majewski J, Zawadzki P, Pickerill P, Cohan FM, & Dowson CG (2000) Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. *J. Bacteriol.* 182(4):1016-1023.

17. Fraser C, Hanage WP, & Spratt BG (2005) Neutral microepidemic evolution of bacterial pathogens. *Proc. Natl. Acad. Sci. U. S. A.* 102(6):1968-1973.

18. Hanage WP, Fraser C, & Spratt BG (2006) Sequences, sequence clusters and bacterial species. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* 361(1475):1917-1927.

19. Falush D*, et al.* (2006) Mismatch induced speciation in *Salmonella*: model and data. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.*361(1475):2045-2053.

20. Maiden MC*, et al.* (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U. S. A.* 95(6):3140-3145.

21. Maiden MC (2006) Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* 60:561-588.

22. Jolley KA, Wilson DJ, Kriz P, McVean G, & Maiden MC (2005) The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. *Mol. Biol. Evol.* 22(3):562-569.

23. Falush D*, et al.* (2001) Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc. Natl. Acad. Sci. U. S. A.* 98(26):15056-15061.

24. Perez-Losada M*, et al.* (2006) Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect. Genet. Evol.* 6(2):97-112.

25. Didelot X & Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175(3):1251-1266.

26. Papke RT*, et al.* (2007) Searching for species in haloarchaea. *Proc. Natl. Acad. Sci. U. S. A.* 104(35):14092-14097.

27. Whitaker RJ, Grogan DW, & Taylor JW (2003) Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science* 301(5635):976-978.

28. Whitaker RJ, Grogan DW, & Taylor JW (2005) Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. *Mol. Biol. Evol.* 22(12):2354-2361.

29. Foti M*, et al.* (2006) Genetic diversity and biogeography of haloalkaliphilic sulphur-oxidizing bacteria belonging to the genus *Thioalkalivibrio*. *FEMS Microbiol. Ecol.* 56(1):95-101.

30. Papke RT, Koenig JE, Rodriguez-Valera F, & Doolittle WF (2004) Frequent recombination in a saltern population of *Halorubrum*. *Science* 306(5703):1928-1929.

31. Boucher Y, Douady CJ, Sharma AK, Kamekura M, & Doolittle WF (2004) Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. *J. Bacteriol.* 186(12):3980-3990.

32. Davelos AL, Xiao K, Samac DA, Martin AP, & Kinkel LL (2004) Spatial variation in *Streptomyces* genetic composition and diversity in a prairie soil. *Microb. Ecol.* 48(4):601-612.

33. Antony-Babu S & Goodfellow M (2008) Biosystematics of alkaliphilic streptomycetes isolated from seven locations across a beach and dune sand system. *Antonie van Leeuwenhoek* 94(4):581-591.

34. Wawrik B*, et al.* (2007) Biogeography of actinomycete communities and type II polyketide synthase genes in soils collected in New Jersey and Central Asia. *Appl. Environ. Microbiol.* 73(9):2982-2989.

35. Lindholm P*, et al.* (1997) *Streptomyces* spp. Isolated from Potato Scab Lesions Under Nordic Conditions in Finland. *Plant Disease* 81(11):1317-1322.

36. Wanner LA (2006) A Survey of Genetic Variation in *Streptomyces* Isolates

Causing Potato Common Scab in the United States. *Phytopathology* 96(12):1363-1371.

37. St-Onge R, Goyer C, Coffin R, & Filion M (2008) Genetic diversity of *Streptomyces* spp. causing common scab of potato in eastern Canada. *Syst. Appl. Microbiol.* 31(6-8):474-484.

38. Flores-Gonzalez R, Velasco I, & Montes F (2008) Detection and characterization of *Streptomyces* causing potato common scab in Western Europe. *Plant Pathology* 57(1):162.

39. Vos M & Velicer GJ (2008) Isolation by distance in the spore-forming soil bacterium *Myxococcus xanthus*. *Curr. Biol.* 18(5):386-391.

40. Hopwood DA (2006) Soil to genomics: The *Streptomyces* chromosome. *Annu. Rev. Genet.* 40:1-23.

41. Flardh K (2003) Growth polarity and cell division in *Streptomyces*. *Curr. Opin. Microbiol.* 6(6):564-571.

42. Jakimowicz D*, et al.* (2000) Architecture of the *Streptomyces lividans* DnaA protein-replication origin complexes. *J. Mol. Biol.* 298(3):351-364.

43. Ohnishi Y*, et al.* (2008) Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. *J. Bacteriol.* 190(11):4050-4060.

44. Chen CW, Huang CH, Lee HH, Tsai HH, & Kirby R (2002) Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. *Trends Genet. : TIG* 18(10):522-529.

45. Bentley SD*, et al.* (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417(6885):141-147.

46. Donadio S, Sosio M, & Lancini G (2002) Impact of the first *Streptomyces* genome sequence on the discovery and production of bioactive substances. *Appl. Microbiol. Biotechnol.* 60(4):377-380.

47. McAlpine JB*, et al.* (2005) Microbial genomics as a guide to drug discovery and structural elucidation: ECO-02301, a novel antifungal agent, as an example. *J. Nat. Prod.* 68(4):493-496.

48. Donadio S*, et al.* (2002) Microbial technologies for the discovery of novel bioactive metabolites. *J. Biotechnol.* 99(3):187-198.

CHAPTER 2

A NEUTRAL MODEL FOR THE EFFECT OF HOMOLOGOUS RECOMBINATION ON MICROBIAL DIVERSIFICATION

Introduction

Widespread evidence of horizontal gene transfer (HGT) among bacteria and archaea has provoked ongoing debate as to whether genetic clusters in these organisms represent species or simply represent points along a continuum of genetic exchange. A variety of evolutionary forces have been invoked in these debates, but chief among them are periodic selection, random drift, and the effects of both HR and non-homologous recombination (NR). Several different models have been proposed to explain the evolution of microbial species (1). The ecotype model, for example, predicts that periodic selection and genetic drift combine to create clonal genetic clusters that each correspond to a unique ecological niche (2). The ecotype model explains well the patterns of divergence observed in *Bacillus spp*. from Evolution Canyon, Israel, where distinct genetic clusters occur in ecologically distinct locations within the canyon (3). In contrast, organisms like *Neisseria meningitidis*, *Streptococcus pneumoniae* and *Helicobacter pylori* have high rates of gene exchange (4, 5). The application of the ecotype model to these populations is complicated by the reshuffling of alleles between ecotypes made possible by HR. While HGT can act as a diversifying force when it involves the acquisition of new genes and traits, HGT can also act as a cohesive force by mediating homologous replacement of existing genes (as reviewed in (6)). HR is the nonreciprocal, unidirectional transfer of a homologous segment of DNA from a donor to a recipient (analogous to interallelic gene conversion in eukaryotes). Since this form of HR is nonreciprocal, it decreases nucleotide

divergence between donor and recipient, with the net change in similarity a function of the sequence divergence prior to replacement.

Methods

We developed a new population simulator that we call *bactsimDF* (for bacterial simulation different and fixed) to model the evolution of bacterial populations forward in time. The *bactsimDF* simulator was developed using the algorithm and code from the forward in time population simulator *forwsim* (7). The algorithm used by *forwsim* to increase simulation speed is detailed elsewhere (7), so we will describe it briefly. The population being modeled is a neutral Wright-Fisher population that is constant in size. Individual chromosomes are modeled as collections of polymorphic sites, allowing new mutations to appear only at non-polymorphic sites. Mutations are tracked by their location on the chromosome. Further information is not necessary, as the simulation works on a pseudo infinite-sites assumption. To improve the speed of the program, *forwsim* simulates the future genealogy of the population in the next 8 generations and tracks all individuals still in the population at that time and any individuals that have recombined with any of those present individuals in that span of time. Mutations or recombination events that occur between individuals that do not contribute to future generations are not simulated. The user may also specify the number of generations after which homogenous features of the population are purged, i.e., if a mutation goes to fixation or extinction, it is no longer necessary to include that polymorphic site in every individual's chromosome and is removed. If a mutation is no longer carried in the population, then it is removed from the list of existing mutations and new mutations may again enter the population at that site (i.e. that location becomes non-polymorphic). The *bactsimDF* simulation program is modified to include homologous recombination (interallelic gene conversion) instead of crossing-over. In addition, the mean tract length is user-specified and for individual events the

tract length is realized as a draw from a geometric distribution, as in the coalescent simulation ms (8). The probability of homologous recombination for each individual per-generation is the specified homologous recombination rate. The sequence divergence (π) between donor and recipient is used to calculate a modified homologous recombination rate between donor and recipient as $R = \rho 10^{-d\pi}$, where ρ is determined by the user defined homologous recombination rate and d is the distance factor, also defined by the user. User inputs include population size, sample (output) size, genome length, total number of generations, the number of generations between deletions of homogenous features, recombination rate, mutation rate, tract length, and distance factor.

The *bactsimDF* program simulates two independent but co-existing populations that start at a user-specified level of nucleotide divergence, and population size. These independent populations are assumed to be in physical contact (allowing genetic transfer) but are otherwise distinct and isolated from the consequences of genetic drift in the neighboring population. Individuals have equal access to DNA from both populations without respect to population boundaries (at a rate modified by the level of nucleotide divergence), but after each round of recombination the new generation of each population is sampled at random from the previous members of the same population only. The purpose of this program is to determine whether populations will converge or diverge based on recombination, mutation, nucleotide divergence and the distance factor. This framework is ideal for simulating ecologically distinct populations that fill different niche spaces, as in the ecotype model (2). The program calculates nucleotide divergence between populations every 500 generations by comparing the user-specified sample size number of individuals in both populations. The parameters used for all of the simulations were: number of individuals, 2,000

(1,000 in each population); sample size, 200; total generations, 25,000,000; mutation rate, 5e-5 individual^{-1} generation^{-1}; recombination rate, variable; time between deletions, 50 generations; tract length, 500; distance factor, -20. *bactsimDF* source code and precompiled binaries are freely available at https://sites.google.com/site/doroghazi/.

To determine the relative change in each population when simulating populations of different sizes, as opposed to the overall change in nucleotide divergence, every 500 generations each population was compared to a sample of its ancestors 500 generations in the past, and to a sample of the other population 500 generations in the past. This data was used to attribute a proportion of the overall change in nucleotide diversity to each population.

The data sets and ClonalFrame output files from Vos and Didelot were graciously provided by Dr. Michiel Vos. Nucleotide diversity was calculated with a Perl script. Values for μ/ρ and π are provided in Table 2.1.

Results and Discussion

Theory

The impact of HR on the divergence of two sympatric populations can be understood as a function of their initial nucleotide divergence and the ratio of mutation to recombination. The effect of mutation on these populations is straightforward, as mutations accumulate in one lineage at the mutation rate of individuals in the population (9), so two isolated populations diverge at twice their mutation rate, or 2μ. The HR rate for any given fragment of DNA is expected to vary as a log-linear function of the nucleotide divergence between donor and recipient. The effect of

sequence similarity on HR rate has been quantified in *Escherichia coli* (10), *Streptococcus pneumoniae* (11), *Bacillus subtilis*, and *Bacillus mojavensis* (12) as roughly $R = \rho 10^{-20\pi}$, where R is the modified recombination rate, ρ is the recombination rate when there are no differences between sequences, π is nucleotide divergence between the sequences involved, and -20 is the slope of the log-linear line that describes the relationship between R, ρ and π (6). Two recombining populations can then be expected to converge due to HR at the rate $2\rho\pi 10^{-20\pi}$ where ρ is the interpopulation recombination rate. The forces of mutation and recombination between two populations are equal when $2\mu = 2\rho\pi 10^{-20\pi}$, or $\mu/\rho = \pi 10^{-20\pi}$. Thus, we can predict that under neutral conditions two populations will diverge when $\mu/\rho > \pi 10^{-20\pi}$ and converge when $\mu/\rho < \pi 10^{-20\pi}$. The change in nucleotide divergence per generation is calculated as $\Delta = 2\mu - 2\rho\pi 10^{-20\pi}$ (Figure 2.1). It is important to note that ρ is the per-site frequency of recombination, combining the rate of occurrence of recombination events and the tract lengths of those events. Populations with interpopulation values of μ/ρ and π that fall below the curve will be subject to the cohesive effects of recombination, manifested as a decline in interpopulation nucleotide divergence, while populations above the curve will be primarily clonal, escaping cohesion and diverging through accumulation of mutations. While we have formulated this theory within the ecotype framework, these equations would also apply to genetic clusters within the same population. However, as the two groups would no longer be independently drifting, random drift would be able to drive clusters to extinction, resetting the level genetic diversity within the population regardless of how distinct the clusters have become.

Forward in time simulations

A novel population simulator, *bactsimDF*, was developed to test the ability of our null

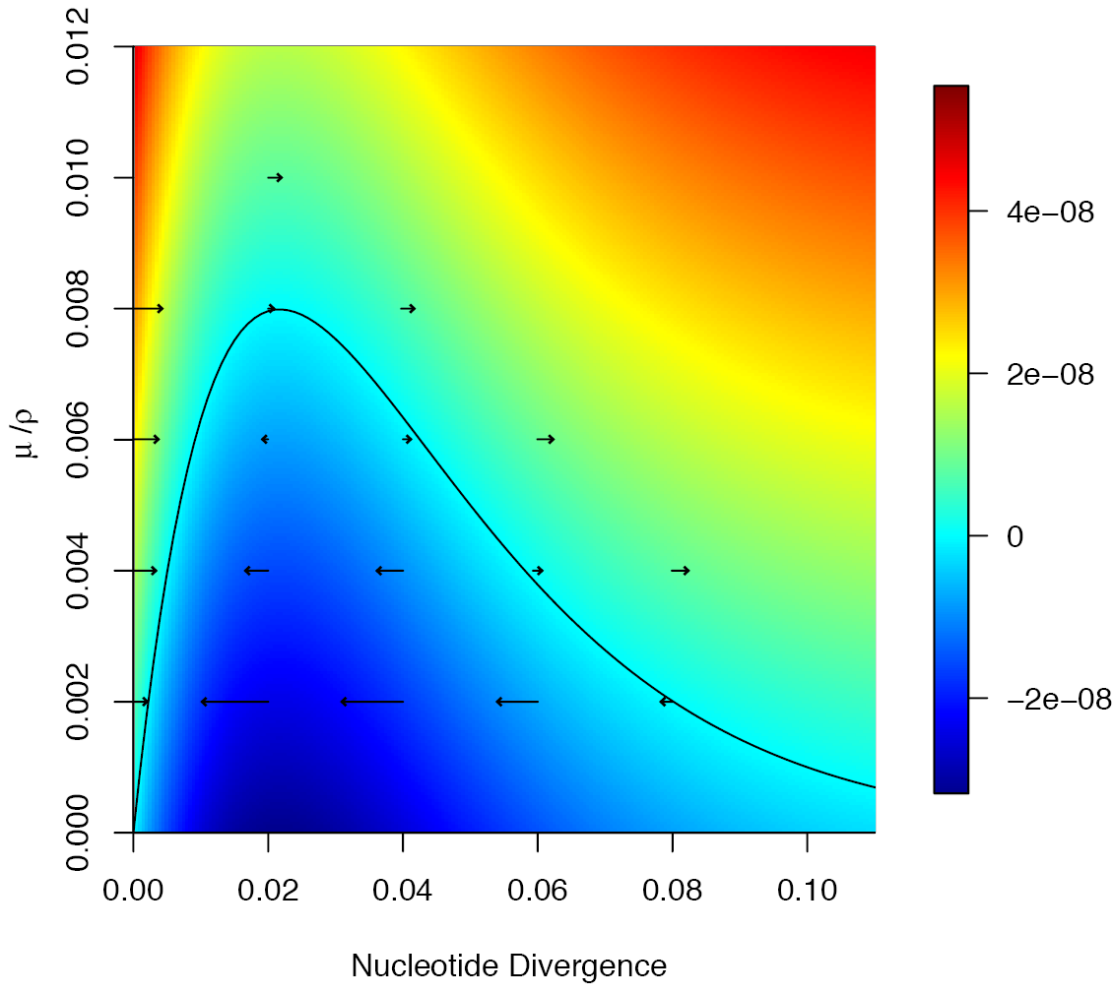


Figure 2.1. Expected and observed relationships between the HR rate and nucleotide divergence between independent populations that are free to recombine. The curve ($\mu/\rho = \pi 10^{-20\pi}$) represents the expected equilibrium that bounds populations subject to the cohesive effects of HR. Lineage convergence due to HR is expected below the curve while lineages above the curve are free to diverge. The heat map represents the expected rate of change in nucleotide divergence between two populations per generation, $\Delta = 2\mu - 2\rho\pi 10^{-20\pi}$. Arrows indicate the observed results from *bactsimDF* simulations run at 5 different values for μ/ρ. The origin of each arrow shows the starting nucleotide divergence, and then end shows the final nucleotide divergence.

model to estimate the evolutionary trajectory of populations as a function of interpopulation nucleotide divergence and recombination rate (Figure 2.1). The simulation program (developed by modifying the population simulator *forwsim* (7) as described in the Methods) makes neutral assumptions to simulate the evolution of bacterial populations forward in time, making it possible to examine the impact of HR and other population parameters on the evolution of bacterial populations. The simulation starts with two populations of defined nucleotide divergence that are free to recombine. Homologous recombination occurs between pairs of individuals selected at random without regard to population of origin, and the success of each event depends upon the sequence similarity of donor and recipient over the tract to be transferred and the user-defined distance factor. Future generations consist of fixed numbers of individuals sampled randomly from each population, allowing drift to act independently on each population. The simulation is ideal for evaluating the impact of gene exchange on sympatric populations that fill different ecological niches. The simulator is useful for testing the impact of recombination on the ecotype model as lineage divergence in the ecotype model requires a founder subpopulation to escape from periodic selection due to acquisition of adaptive alleles. The ability of the model to simulate two coexisting populations that are able to recombine but are otherwise ecologically independent from each other mirrors the events that would be expected when a new ecotype forms in sympatry. Simulations consisted of two populations of 1,000 individuals each, having genomes of 500,000 bases, and mean recombination tract length of 500 bases. The change in nucleotide divergence that occurred during simulations agrees well with the calculated expected change in nucleotide divergence (Figure 2.2). Populations with simulated parameters that fall above the curve are observed to diverge while those below the curve converge (Figure 2.1), with one exception. A single simulation started to oscillate instead of converging (Figure 2.1,

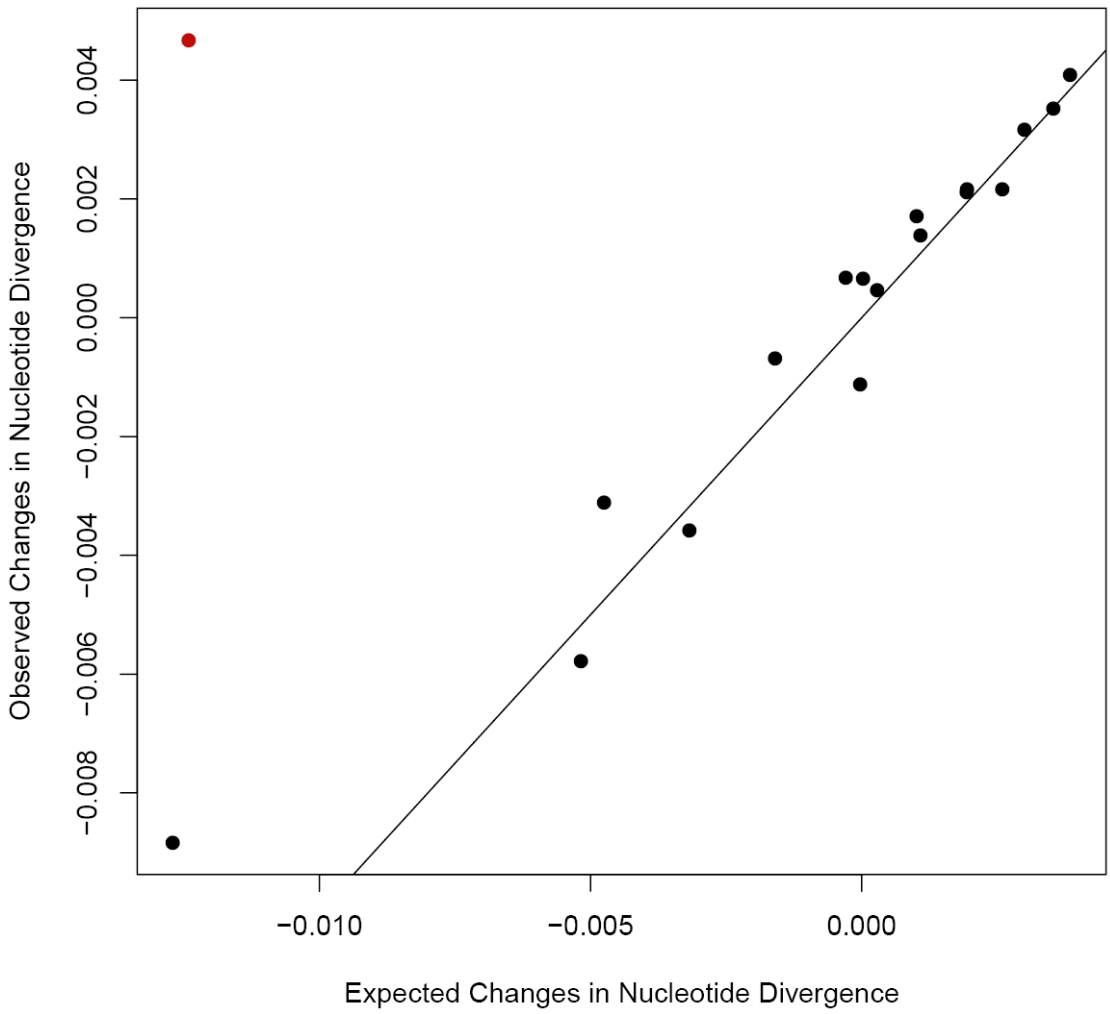


Figure 2.2. Observed changes in nucleotide divergence in simulations over 25 million generations compared to expected changes calculated using recurrent application of the formula $\Delta = 2\mu - 2\rho\pi 10^{-20\pi}$. The line shown has an intercept of 0 and a slope of 1; identical observations and expectations should fall exactly on this line. The outlier, shown in red, is the one simulation that diverged and oscillated instead of converging (see Figure 2.1).

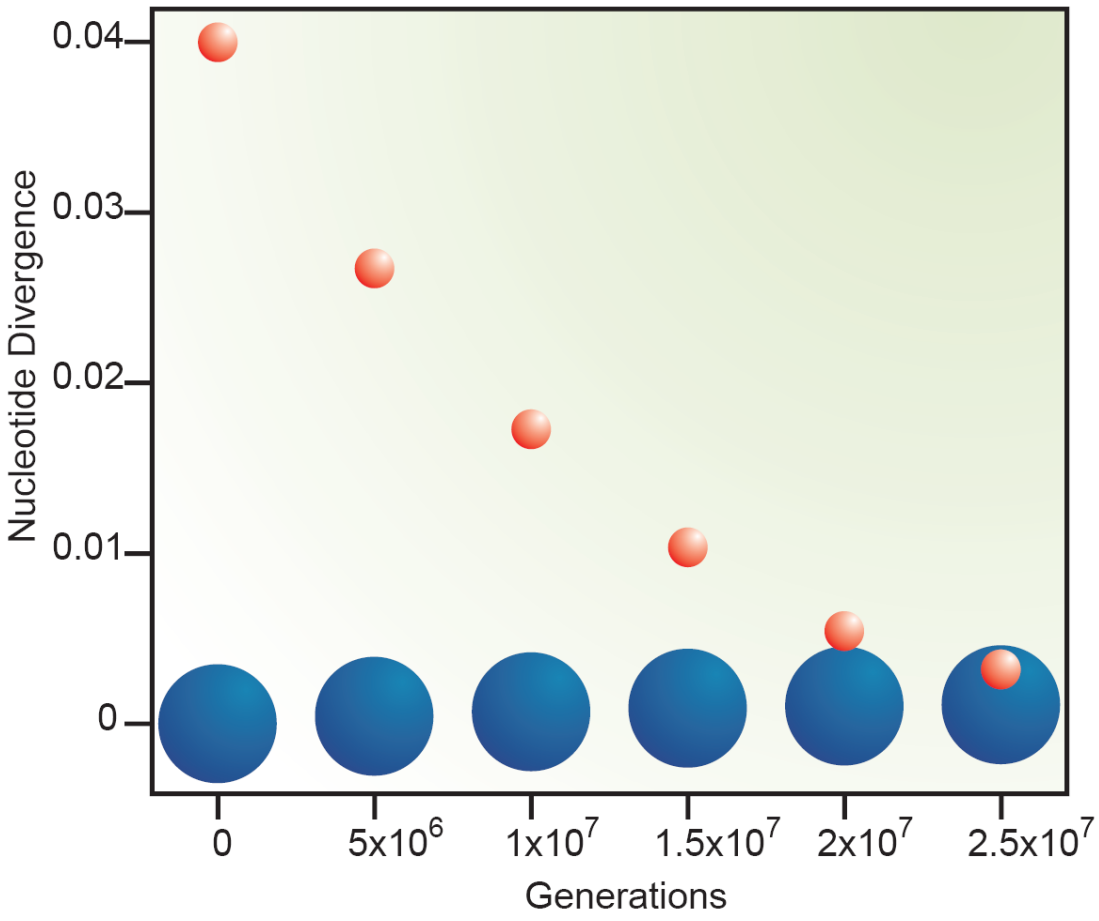


Figure 2.3. Change in nucleotide divergence in two recombining populations that differ in size. Circle area is proportional to population size with 1800 individuals in the large population and 200 in the small population. The simulation was started at nucleotide divergence of 4% with μ/ρ of 0.00035.

Figure 2.2), most likely due to stochastic variation caused by the proximity of the population parameters used in this particular simulation to an unstable equilibrium between the forces of recombination and mutation.

The model we describe can be extended to evaluate recombination between populations that differ in size by accounting for the change in interpopulation recombination rate due to the different probability of encountering DNA from an individual from the other population (this relates to propinquity and not necessarily to differences in effective population size). For equal sized populations, $\rho_1\pi10^{-20\pi} + \rho_2\pi10^{-20\pi} = 2\rho_1\pi10^{-20\pi}$ as $\rho_1 = \rho_2$, but for populations that differ in size the individual interspecies recombination rate must be calculated for each population. To simulate coexisting populations of unequal size, we modified *bactsimDF* to include one population of 1800 individuals and one of 200 individuals (Figure 2.3). When averaged across all individuals in both populations the recombination rate for populations of unequal size and for those of equal size was the same (5×10^{-4} individual^{-1} generation^{-1} with a tract length of 500), but when examined with respect to the different populations it was determined that the smaller population realized an interpopulation recombination rate of 4.74×10^{-4}, 5.2 times higher than that realized for the larger population. As a result, the smaller population is swamped with DNA from the larger population, drifting 33.2 times closer to the large population than the large moved towards the small (Figure 2.3). This result has important implications for the ability of ecotype formation to promote the sympatric divergence of lineages. The period during which a nascent ecotype is formed from an ancestral population would, by necessity, be characterized by inequality in population size with the new ecotype representing the minority population.

Relevance to the study of named species

We examined the biological relevance of our model by using values of μ/ρ and π calculated from data reported by Vos and Didelot (13) and provided by Dr. Michiel Vos for a range of bacterial and archaeal species (Figure 2.4). We also examined the impact of varying the distance factor of recombination by 50% (Figure 2.4). The population parameters for a wide range of microbial species can be seen to span the equilibrium line that separates recombining and clonal populations. While these data were not specifically collected with respect to intraspecies population structure, these values provide a useful proxy to assess the ability of a newly formed subpopulation to diverge from its parent species in the absence of barriers to gene exchange. Our model locates the species *Bacillus thuringiensis*, *Bacillus weihenstephanensis*, and *Bacillus cereus* above the equilibrium line regardless of the value of the distance factor (Figure 2.4). The model predicts that species with these parameters will be clonal having subpopulations able to diverge on separate evolutionary trajectories as described by the ecotype model. In contrast, *Helicobacter pylori*, *Neisseria meningitidis*, and *Streptococcus pneumoniae*, which are known to be highly recombining (4, 5), fall under the equilibrium line regardless of the distance factor. The model predicts that the evolution of populations with these parameters would be dominated by the cohesive effects of recombination. While the evolutionary significance of HR differs dramatically for the species described above we can see that these differences arise as manifestations of simple underlying biological principles.

Implications and further discussion

The *bactsimDF* population simulator was designed to test the impact of recombination on sympatric populations occupying different ecological niches. The ecotype model predicts that periodic selection acts to isolate ecotypes into distinct genetic clusters.

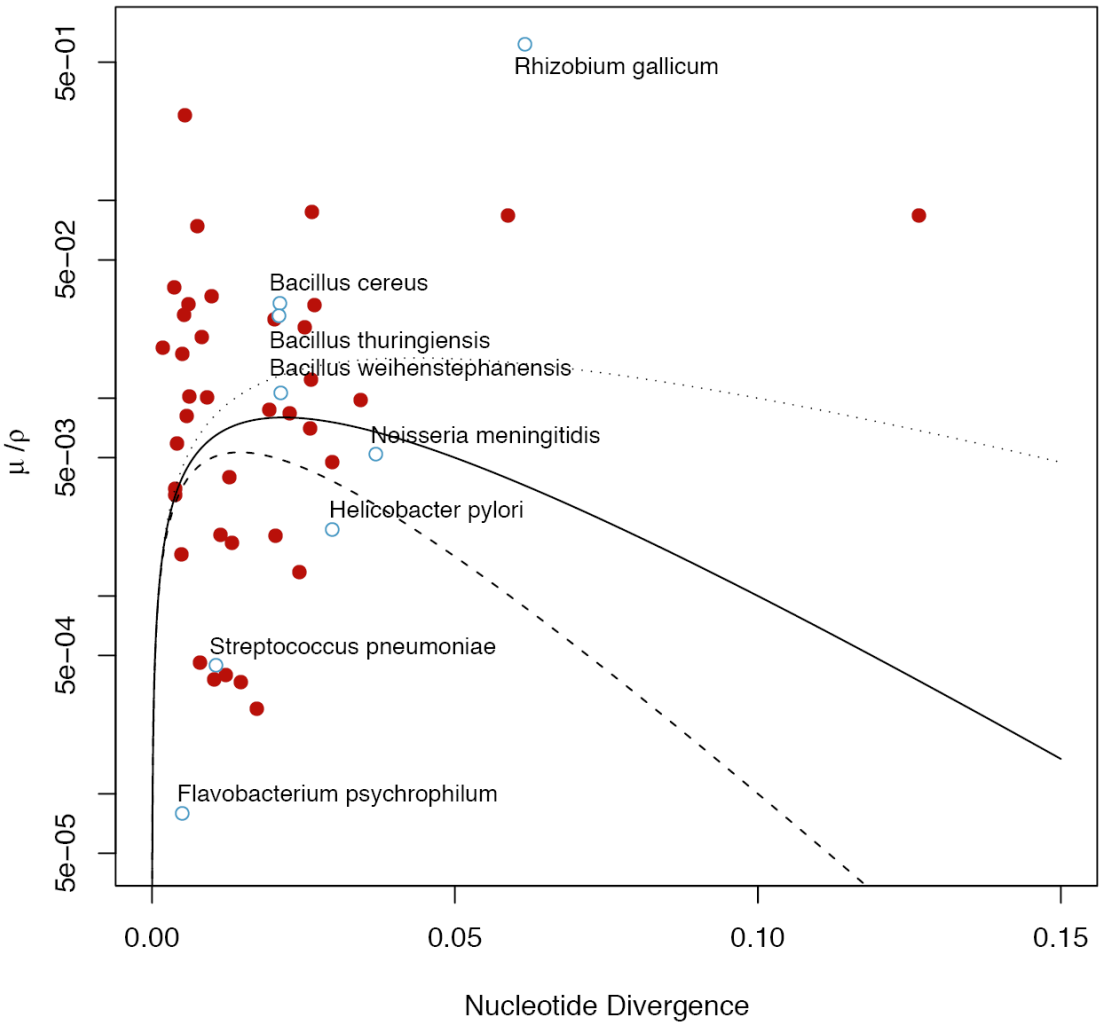


Figure 2.4. Population parameters estimated from characterized microbial species in relation to the equilibrium threshold for recombining populations. In the figure the recombination distance factor was varied to demonstrate the impact of this value on the equilibrium threshold. The distance factors that were used are: -20 for the solid line (as presented in Figure 2.1), –10 for the dotted line, and –30 for the dashed line. Values for μ/ρ are calculated as $\pi(r/m)^{-1}$ from the values of r/m in Vos and Didelot (13) and π is calculated from the same data sets provided by Dr. Michiel Vos.

Under the ecotype model speciation occurs as a consequence of niche invasion and new ecotype foundation. We show that under neutral assumptions, the HR rates estimated for a variety of microbial species would preclude genetic isolation through the mechanism of ecotype foundation. Populations under the equilibrium line (Figure 2.1) would be sufficiently cohesive such that nascent ecotypes would be unable to diverge across the homologous loci of the core genome. We might speculate that in recombining populations niche expanding mutations or gene acquisitions would be found in the auxiliary genome and their exchange within the population governed by propinquity. Application of the ecotype concept would still have potential utility for characterizing the divergence of populations found above the equilibrium line predicted by our theoretical model, such as those of *Bacillus* species.

It is interesting to note that equilibrium threshold for recombining populations (Figure 2.1) declines rapidly as nucleotide identity drops from 98% to 92%. Microbial species established using current taxonomic criteria can be roughly delineated by a 95% genome ANI (average nucleotide identity) criterion (15, (14). Thus, our model provides a theoretical explanation for why the 95% ANI criterion would approximately delineate species that were described using the traditional species concept for bacteria. This prediction from our model may explain the valley of genetic discontinuity that is observed in metagenome surveys of ocean populations and in genomic comparisons, and which has been presented as further evidence for the validity of the 95% ANI criterion for describing microbial species (15). The model we present suggests, however, that a 95% ANI is a blunt tool and that coherent genetic clusters could form over a wide range of ANI values depending on the level of HR in the lineage and the evolutionary mechanisms that have driven cluster formation. As an example, clonal populations that have diverged as a consequence of ecotype

foundation may form genetically and ecologically coherent clusters that might reasonably be considered different species despite having ANI greater than 95%. In contrast, it is possible that two isolates that have 92% ANI may both belong to the same cohesive genetic cluster when rates of recombination are very high (Figure 2.1).

While we have restricted our analyses to the effects of HR, NR also serves as a powerful force of evolutionary change. NR contrasts with HR in that the former will lead to gene acquisition rather than gene replacement, and as a result we would expect that NR will generally manifest as a diversifying rather than as a cohesive force and have its greatest impacts on the auxillary genome. In addition, alleles acquired by HR are not likely to facilitate niche expansion while there are many examples of genes acquired by NR that have adaptive significance and lead to the colonization of new niches (eg: pathogenecity islands). Niche expansion by NR mediated acquisition of adaptive genes can promote ecotype foundation for populations above the equilibrium threshold of recombination, but our model predicts that the acquisition of niche expanding genes would not be expected to promote sympatric divergence within a cohesively recombining population unless somehow the acquisition also created a barrier to gene exchange. In this case it would be the barrier to gene exchange, rather than escape from periodic selection, that would be expected to drive lineage divergence. It is important to consider that NR can also have downstream implications for HR rates. NR results in the insertion of new stretches of DNA that lower the local sequence similarity at the conserved boundary sequences. The local decrease in sequence similarity has the effect of lowering local HR rates and this might result in propagating fronts of divergence as described by Vetsigian *et al*. (16) The propagating front hypothesis provides a mechanism that could promote divergence in highly recombining populations by creating barriers to gene exchange. Further research on

the interplay of both types of recombination is necessary and in the future it should be possible to modify *bactsimDF* to simulate the effects of both HR and NR on the core and auxiliary genome and to evaluate the role of propagating fronts in lineage diversification.

It is interesting to consider the model we present in relation to the species *Campylobacter jejuni* and *C. coli.* Exchange of housekeeping genes between these species has been used as evidence to suggest that these species are merging to form a new lineage as a consequence of interspecies gene exchange (17). However, population parameters for *C. jejuni* fall above the equilibrium line for cohesive recombination as predicted by our model (Table 2.1, Figure 2.4). We assume that interspecies recombination rates will generally be lower than intraspecies recombination rates. Based on this assumption we can predict from the model that these species should be on separate evolutionary trajectories. Merger of these species might occur if favored in some way by selection, but based on neutral assumptions the cohesive force of homologous recombination should be insufficient to drive merger these species. Recent genomic analysis of *C. jejuni* and *C. coli.* reveals that the loci used by Sheppard et al. to estimate recombination rate were influenced by hitchhiking with genes under positive selection (18). As a result, it now seems likely that the recombination rate between *C. jejuni* and *C. coli* was overestimated and that these species are unlikely to merge due to interspecies homologous recombination (18). The availability of a null model built on a theoretical foundation has great utility when investigating bacterial population dynamics as it provides some expectations as to how populations should behave under simplifying assumptions, providing fertile grounds for generating and testing hypotheses about the evolutionary dynamics of specific populations.

Table 2.1. Estimated values of μ/ρ and calculated values for π from the same data sets from Michiel Vos (13). μ/ρ was calculated as $\pi(r/m)^{-1}$.

Sampled population	**π**	**μ/ρ**
Bacillus cereus	0.02110	0.03014
Bacillus thuringiensis	0.02090	0.02613
Bacillus wiehenstephanensis	0.02124	0.01062
Bartonella henselae	0.00181	0.01809
Bordetella pertussis	0.00600	0.03000
Campylobacter jejuni	0.01938	0.00881
Campylobacter insulaenigrae	0.01280	0.00400
Chlamydia trachomatis	0.02645	0.08817
Clostridium difficile	0.00530	0.02650
Enterococcus faecalis	0.00613	0.01022
Enterococcus faecium	0.00385	0.00350
Escherichia coli ET-1 group	0.00416	0.00594
Flavobacterium psychrophilum	0.00499	0.00008
Haemophilus influenzae	0.02610	0.00705
Haemophilus parasuis	0.02266	0.00839
Halorubrum sp.	0.02630	0.01252
Helicobacter pylori	0.02970	0.00218
Klebsiella pneumoniae	0.00984	0.03280
Lactobacillus casei	0.00364	0.03640
Legionella pneumophila	0.00909	0.01010
Leptospira interrogans	0.00539	0.26950
Listeria monocytogenes	0.05870	0.08386
Mastigocladus laminosus	0.02681	0.02978
Microcoleus chthonoplastes	0.02017	0.02521
Microcystis aeruginosa	0.02430	0.00133
Moraxella catarrhalis	0.02040	0.00202
Mycoplasma hyopneumoniae	0.00490	0.00163
Myxococcus xanthus	0.01130	0.00205
Neisseria lactamica	0.02970	0.00479
Neisseria meningitidis	0.03700	0.00521
Oenococcus oeni	0.00570	0.00814
Pelagibacter ubique	0.01730	0.00027
Plesiomonas shigelloides	0.01320	0.00186
Porphyromonas gingivalis	0.00820	0.02050
Pseudomonas syringae	0.12656	0.08437
Ralstonia solanacearum	0.02520	0.02291
Rhizobium gallicum	0.06160	0.61600
Salmonella enterica	0.01220	0.00040
Staphylococcus aureus	0.00742	0.07420
Streptococcus pneumoniae	0.01050	0.00045
Streptococcus pyogenes	0.00788	0.00046
Sulfolobus islandicus	0.00390	0.00325
Vibrio parahaemolytics	0.01460	0.00037
Vibrio vulnificus	0.01020	0.00038
Wolbachia b complex	0.03442	0.00984
Yersinia pseudotuberculosis	0.00505	0.01683

While the neutral theory of evolution has waned in popularity, the utility of neutral theories as null hypotheses for the study of evolution is without question. The model we propose provides criteria for evaluating the evolutionary forces that lead to speciation and the persistence of coherent genetic clusters. In populations above the threshold ratio of mutation to recombination, the ecotype model can explain well the formation of genetic clusters. Efforts to understand speciation in these lineages may require characterization of ecological niches and adaptive alleles. In contrast, in populations found below the threshold ratio of mutation to recombination, multiple ecotypes may be found within one cohesively recombining group. Efforts to understand speciation for these lineages may require characterization of barriers to gene exchange. These barriers may include mating specificity factors (e.g. plasmids), fitness cost associated gene exchange, breakdowns in local sequence similarity due to non-homologous recombination events (e.g. those leading to propagating fronts of divergence (16)), or factors that effect propinquity (e.g. spatial isolation). For example, if two populations with high intraspecies recombination rates were found to co-exist as distinct lineages despite high levels of interpopulation nucleotide similarity the model would generate the hypothesis that some barrier to gene exchange must exist currently and on evolutionary timescales, and would promote efforts to identify such barriers. The model we propose can serve as a foundation for making hypotheses about microbial speciation and as a step towards understanding the variety of evolutionary forces that can generate coherent genetic clusters in microbial lineages.

REFERENCES

1. Fraser C, Alm EJ, Polz MF, Spratt BG, & Hanage WP (2009) The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity. *Science* 323(5915):741-746.
2. Cohan FM (2002) What are bacterial species? *Annu. Rev. Microbiol.* 56:457-487.
3. Cohan FM & Perry EB (2007) A systematics for discovering the fundamental units of bacterial diversity. *Curr. Biol.* 17(10):R373-386.
4. Feil EJ, Enright MC, & Spratt BG (2000) Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between *Neisseria meningitidis* and *Streptococcus pneumoniae*. *Res. Microbiol.* 151(6):465-469.
5. Falush D*, et al.* (2001) Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc. Natl. Acad. Sci. U. S. A.* 98(26):15056-15061.
6. Fraser C, Hanage WP, & Spratt BG (2007) Recombination and the nature of bacterial speciation. *Science* 315(5811):476-480.
7. Padhukasahasram B, Marjoram P, Wall JD, Bustamante CD, & Nordborg M (2008) Exploring population genetic models with recombination using efficient forward-time simulations. *Genetics* 178(4):2417-2427.
8. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337-338.
9. Kimura M (1968) Evolutionary Rate at the Molecular Level. *Nature* 217(5129):624-626.
10. Vulic M, Dionisio F, Taddei F, & Radman M (1997) Molecular keys to

speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc. Natl. Acad. Sci. U. S. A.* 94(18):9763-9767.

11. Majewski J, Zawadzki P, Pickerill P, Cohan FM, & Dowson CG (2000) Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. *J. Bacteriol.* 182(4):1016-1023.
12. Zawadzki P, Roberts MS, & Cohan FM (1995) The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust. *Genetics* 140(3):917-932.
13. Vos M & Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 3(2):199-208.
14. Konstantinidis KT, Ramette A, & Tiedje JM (2006) The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 361(1475):1929-1940.
15. Konstantinidis KT & DeLong EF (2008) Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J.* 2(10):1052-1065.
16. Vetsigian K & Goldenfeld N (2005) Global divergence of microbial genome sequences mediated by propagating fronts. *Proc. Natl. Acad. Sci. U. S. A.* 102(20):7332-7337.
17. Sheppard SK, McCarthy ND, Falush D, & Maiden MC (2008) Convergence of *Campylobacter* species: implications for bacterial evolution. *Science* 320(5873):237-239.
18. Caro-Quintero A, Rodriguez-Castano GP, & Konstantinidis KT (2009) Genomic Insights into the Convergence and Pathogenicity Factors of *Campylobacter jejuni* and *Campylobacter coli* Species. *J. Bacteriol.* 191(18):5824-5831.

CHAPTER 3

WIDESPREAD HOMOLOGOUS RECOMBINATION WITHIN AND BETWEEN *STREPTOMYCES* SPECIES

Introduction

Evidence for extensive genomic heterogeneity and widespread gene exchange in microbial genomes has highlighted our limited understanding of microbial species, and caused the very idea of species to be questioned (1). Several models have been proposed to explain the origin and persistence of genetic clusters that correspond to microbial species (as reviewed (2)). Ecotype models, which predict that adaptive mutations define a species as an occupant of a distinct ecological niche, are theoretically appealing and work well for some taxa (3). Even in environments where the ability of periodic selection to purge diversity is limited by the patchy distribution of habitable sites in space and time, associations between genetic diversity and environmental specialization have been observed (4). Ecotype based models may be less appealing, however, for microbial populations characterized by high rates of gene exchange in which recombination limits the ability of periodic selection to purge genetic diversity (2). In such highly recombining populations, the origin and persistence of species may be defined primarily by biological, ecological, or geographical barriers to gene exchange (2, 5). For example, the recent elimination of ecological barriers to gene exchange has been suggested as a mechanism to explain patterns of admixture resulting from hybridization between *Campylobacter* species (6). The general significance of homologous gene exchange in defining microbial species remains unclear, however, as rates of homologous exchange can vary widely, and available data mainly focus on pathogenic populations which represent a narrow

spectrum of the ecological and phylogenetic diversity in the microbial world (7).

Streptomycetes represent an interesting model system for exploring the impact of gene exchange on the genetic diversity of microbial taxa. Originally classified as fungi, streptomycetes are Gram positive bacteria which have a complex life cycle characterized by hyphal growth and mycelium formation followed by development of aerial hyphae and asexual production of spores. *Streptomyces* species have a large (≥ 7-9Mb) linear chromosome, which contains a central region that is highly conserved throughout the genus, and terminal regions that are variable in composition and organization (8). Conjugation in streptomycetes involves transfer of double stranded DNA (dsDNA) and requires the product of only a single gene, *traB*, which encodes a septal DNA translocator protein (9). Gene exchange takes place at the tips of elongating filaments and may be facilitated by hyphal fusion (8). Chromosomal markers are mobilized through this mechanism at a frequency of 0.1 – 1% (10), with plasmid integration causing mobilization of chromosomal genes at an efficiency that approaches 100% (11). This mechanism is in stark contrast to that observed in Gram negative bacteria, which involves transfer of single stranded DNA and requires the coordinated action of multiple proteins and an origin of transfer. The phylogenetic distribution of these two mechanisms of gene exchange among bacteria remains poorly characterized, but the dsDNA exchange system has been most clearly documented in multicellular Gram positive bacteria. Surveys of homologous recombination rates do not currently include any populations which possess this dsDNA conjugation system.

Given the nature of gene exchange and recombination mechanisms in streptomycetes we hypothesized that homologous recombination may have a significant impact on

their evolution. We tested this hypothesis at the intraspecies level using strains from a single *Streptomyces* species. A collection of 37 isolates of *Streptomyces flavogriseus* was obtained from five sites spanning four counties in New York and one in Michigan. MLSA of these isolates was used to estimate homologous recombination rates for the *S. flavogriseus* population. We also analyzed an existing MLSA dataset of *Streptomyces* spp. to quantify rates of interspecies gene exchange and to evaluate the impact of HGT on the evolution of *Streptomyces*.

Materials and Methods

Soil sampling and isolation of S. flavogriseus

S. flavogriseus isolates were obtained from soil samples (0-5 cm depth) taken from grassy fields identified in five locations in New York and Michigan. The sites sampled in New York were: the Miner Institute, Chazy, Clinton County (N 44.884672, W -73.474429); Willsboro Farm, Willsboro, Washington County (N 44.385817, W -73.384850); Mitchell Street Field, Ithaca, Tompkins County (N 42.434654, W -76.471442); Caldwell Field, Ithaca, Tompkins County (N 42.450061, W -76.458782); and Harford Farm, Harford, Courtland County. The Michigan site was in Kentwood, Kent County (N 42.856419, W -85.622737). Fresh soil was diluted 1:100 in phosphate buffered saline solution and 50 µL of this suspension was spread onto glycerol-arginine agar plates with a pH of 8.5-8.7 containing 300 mg l^{-1} cycloheximide (12). The most common streptomycete colony type on this media was white on the edges and become dark gray in the center after 5-7 days, at which point they are 1-2 mm diameter. The 16S rRNA and *rpoB* gene sequences of each isolate were screened (see MLSA below) and the 37 isolates obtained in this manner were found to be highly similar (greater than 99.8 % sequence similarity, Table 3.1).

The gene sequences from these 37 isolates were found to be highly similar (99.8% average sequence similarity across the 5 protein coding genes examined, see below) to gene sequences found in the genome of *S. flavogriseus* IAF-45-CD (ATCC 33331, genbank accession: ACZH00000000). *S. flavogriseus* IAF-45-CD was originally isolated in Laval, Canada at a site less than 100 km from Chazy, NY (Ishaque & Kluepfel, 1980). In terms of morphology and carbon-source utilization profile (D-glucose (+), L-arabinose (+), D-xylose (+), raffinose (-), D-fructose (+), I-inositol (-), D-mannitol (+), rhamnose (+)) our isolates match *S. flavogriseus* as described in the International *Streptomyces* Project (13) and in the Wink Compendium (14). It should be noted that *S. flavogriseus* IAF-45-CD (ATCC33331) and the type strain *S. flavogriseus* Heim (ATCC25452) share only 93.5% similarity across the 5 protein coding genes examined and 97.1% 16S rRNA similarity suggesting that our 37 isolates and *S. flavogriseus* IAF-45-CD represent a population that is genetically distinct from *S. flavogriseus* Heim. Collectively we refer to our 37 strains and strain IAF-45-CD as *S. flavogriseus*. phylogroup *pratensis* (from the latin for 'growing in the meadow'). It should also be noted that the gene sequences provided for a strain described as 'S. *griseoplanus'* in Guo et al. (2008) match exactly those from strain IAF-45-CD and given that the 16S rRNA gene provided for 'S. *griseoplanus'* by Guo et al. (2008) does not match the 16S rRNA gene sequence deposited for the type strain of S. *griseoplanus* (ACCN: AB184138.1) it seems likely that strain IAF-45-CD was misidentified as 'S. *griseoplanus'* in Guo et al. (2008) .

MLSA of Streptomyces sp.

DNA was extracted and the MLSA scheme of Guo *et al*. (15) was used to characterize the isolates. Due to reported problems in Guo *et al*. (2008) with existing *gyrB* primers,

we designed new primers using recently released genome data from additional *Streptomyces* species (gyrBF: CTG GAC GCG GTC CGC AAG CG; gyrBR: GTC TGG CCC TCG AAC TGC GGC T). All reactions were performed with the following 25µL reaction using AmpliTaq Gold reagents (Applied Biosystems, NJ, USA): 11.75µl H_2O, 2.5µl 10x Buffer, 3µl 25mM $MgCl_2$, 2µl dNTP mixture (2.5mM each dNTP, 10mM total dNTPs, Promega, WI), 1µl forward primer from 10µM stock, 1µl reverse primer from 10µM stock, 2.5µl DMSO, 0.25µl AmpliTaq Gold (5U/µl), and 1µl template. For all primer sets, the following reaction conditions were used: 95°C, 10 min. for initial denaturation; 35 cycles of 95°C for 20 s, 65°C for 30 s, 72°C for 45 s; 72°C for 10 min. as a final extension; short-term storage at 4°C. Sequences were assembled manually and trace files were inspected for all sequences at all polymorphic sites. To confirm results and verify the absence of cross-contamination, isolates Cald 193, Chazy 277, Harf 495, MS7 19, W25 20, W25 23, W25 25, W25 26, and W300 21 were removed from storage and the sequence of the *rpoB* and *traB* genes were determined for 1-4 different individual colonies from each isolate. In every case the expected sequence type was recovered. Gene sequences are available from Genbank with accession numbers GU979234-GU979418.

Assessment of homologous recombination and population structure

The properties of individual loci used for interspecies and intraspecies comparisons are provided in Table 3.1. Sequences of *Streptomyces* species were acquired from NCBI using accession numbers provided in Guo *et al* (15). The 16S rRNA gene sequences were included in the concatenated alignment used for interspecies analyses, but were not used in intraspecies analysis due to a lack of polymorphism. The standardized index of association was calculated with LIAN v3.5 (16) for allelic data from the *S. flavogriseus* isolates. LDhat (17) was used to estimate ρ both for

concatenated sequences and single loci. ClonalFrame was run with 100,000 burn-in updates followed by 100,000 more updates on data without genome positions included (18). Maximum likelihood (ML) trees were created for each locus from the 53 *Streptomyces* species described in Guo *et al* using PAUP v4.0Beta. Maximum likelihood trees were made using the tree bisection reconnection algorithm and incongruence between trees was then evaluated using the Shimodaira-Hasegawa test (19). Individual recombination events within the interspecies data set were found using the Recombination Detection Program v3b34 (20). Reported *p*-values are calculated using the Bonferroni-correction within the program. Neighbornet phylogenetic networks were created with Splitstree v4.10 (21).

Table 3.1. Properties of the loci used in *Streptomyces* MLSA. Values for Tajima's D were not significant (n.s.). Values for ρ and θ_w were calculated using LDhat (17) and are expressed per site with ρ expressed as the rate of gene conversion, $\rho = 2N_e r/2$. Values of ρ and θ_w could not be determined (n.d.) for certain loci due to the low number of polymorphic sites observed.

	Gene	Length	π	Sites	D	P	ρ	θ_w	ρ/θ_w
Interspecies	16S	1389	0.0162	104	-0.850	n.s.	0.0061	0.0163	0.37
	recA	504	0.0766	168	-0.690	n.s.	0.0704	0.0625	1.13
	atpD	496	0.0746	146	-0.220	n.s.	0.0121	0.0511	0.24
	rpoB	540	0.0832	197	-0.624	n.s.	0.0546	0.0645	0.85
	gyrB	423	0.1179	145	-0.601	n.s.	0.1280	0.0748	1.71
	trpB	571	0.0853	193	-1.030	n.s.	0.2150	0.0567	3.80
Intraspecies	*recA*	504	0.0000	0	n.d.	n.d.	n.d.	n.d.	n.d.
	atpD	496	0.0009	3	-0.849	n.s.	n.d.	0.0014	n.d.
	rpoB	540	0.0022	7	-0.799	n.s.	0.1470	0.0031	47.6
	gyrB	417	0.0017	3	-0.035	n.s.	0.0025	0.0017	1.47
	trpB	571	0.0039	8	0.451	n.s.	0.0092	0.0033	2.76

Given the extent of recombination observed between *Streptomyces* species the program Structure was also used to examine population structure among these taxa. Structure was run on data with 20,000 burn-in and 100,000 updates using the linkage model with other parameters set to default (22). Structure assumes that populations are in both linkage equilibrium and Hardy-Weinberg equilibrium. While there is ample evidence for homologous gene exchange between species in *Streptomyce*s, this evidence is insufficient in itself to confirm that these *Streptomyces* satisfy the assumptions of the Structure analysis. While the linkage model is designed to relax these assumptions and permit clustering for a wide range of population structures (22), it is important to consider the impact of violating these assumptions. The admixed model in Structure tries to find the largest populations that are in equilibrium with admixture introduced to cope with linkage disequilibrium (22). True admixture is generally asymmetrically distributed across individuals. Thus, in the absence of actual structure the default assumption would be that the equal distribution of ancestral populations across individuals (22).

Results and Discussion

Intraspecies homologous recombination

A total of 31 sequence types (ST) were detected in the collection of 38 *S. flavogriseus* isolates (Figure 3.1). Identical alleles were observed in a variety of sampling sites and in a wide variety of combinations that could only result from recombination. In total, 5 sequence types and 15 of the 30 alleles observed were present in two or more sites separated by more than 300 kilometers (Figure 3.1). The standardized index of association for the population (Table 3.2) is one of the lowest ever calculated for a bacterial or archaeal population indicating that the population is in almost perfect linkage equilibrium. This result could only be obtained if *S. flavogriseus* phylogroup

Figure 3.1. Allele information for the *S. flavogriseus* phylogroup *pratensis* isolates. Each allele for each locus is represented by a unique number and color. The upper part of the figure provides information on the polymorphic sites for each allele at each locus, with nucleotide positions corresponding to the concatenated sequence alignment. ST indicates the sequence type; # indicates the number of times each ST was recovered; and site indicates the isolation source with *a* and *b* representing sites in northern NY (in Willsboro and Chazy respectively), *c*, *d*, and *e* representing sites in central NY (Mitchell Street Field and Caldwell Field in Ithaca, and Harford Farm in Harford respectively), *f* representing the site in Michigan, and *g* representing the strain characterized by Guo et al. (2009).

	atpD			*rpoB*							*gyrB*			*trpB*							
Allele	724	745	758	1300	1330	1333	1369	1399	1411	1492	1643	1823	1955	1968	1977	2301	2302	2322	2415	2463	2514
1	C	C	C	G	C	C	T	A	T	C	C	T	C	C	G	C	G	G	T	C	C
2	C	T	C	G	C	C	T	G	C	C	C	C	C	T	C	C	G	A	T	A	C
3	T	C	C	G	C	C	T	G	T	T	T	T	C	C	C	C	G	A	T	A	C
4	T	T	C	G	C	C	T	G	T	C	C	T	A	C	C	C	G	A	A	C	C
5	C	C	T	G	T	T	T	G	T	C				T	C	C	G	G	T	C	T
6				G	C	C	C	G	C	C				T	C	C	G	A	A	C	C
7				G	T	T	T	G	C	C				T	C	C	G	G	T	C	C
8				T	C	C	T	G	C	C				C	C	C	T	A	A	C	C
9														C	C	C	G	A	T	C	C
10														T	C	C	G	A	T	C	C
11														C	C	T	G	A	T	A	C
12														C	C	C	G	A	T	C	T

ST	#	Site	*recA*	*atpD*	*rpoB*	*gyrB*	*trpB*
1	1	d	1	1	1	1	1
2	1	c	1	1	1	2	4
3	4	acdf	1	1	2	1	3
4	1	a	1	1	2	1	4
5	2	ae	1	1	2	1	6
6	1	c	1	1	2	2	3
7	1	c	1	1	2	2	7
8	1	f	1	1	2	3	5
9	1	a	1	1	2	4	12
10	1	f	1	1	3	1	4
11	2	ac	1	1	4	1	3
12	1	a	1	1	4	1	9
13	1	a	1	1	4	1	11
14	2	cf	1	1	4	2	3
15	1	a	1	1	4	3	1
16	1	a	1	1	4	3	10
17	1	f	1	1	5	3	5
18	2	ae	1	1	6	1	1
19	1	a	1	1	6	1	5
20	1	a	1	1	6	1	9
21	1	c	1	1	7	1	3
22	1	a	1	1	7	3	5
23	1	a	1	1	8	2	9
24	1	b	1	2	2	1	2
25	1	a	1	2	4	1	2
26	1	g	1	2	4	3	5
27	1	a	1	3	2	1	3
28	1	a	1	3	2	2	4
29	1	e	1	3	4	2	3
30	1	a	1	4	2	3	8
31	1	a	1	5	2	1	3

Table 3.2. Estimated population parameters. Values for sequence length (Length) and polymorphic sites (Sites) are in nucleotides. The value for nucleotide diversity (π) is the average number of nucleotide differences per site. The value presented from the Φ_w test is the P-value calculated under the null hypothesis of no recombination. The value for the standardized index of association (I_a^s) is 0 for a population in linkage equilibrium. Values for ρ and θ_w were calculated using LDhat (17) and are expressed per site for the concatenated gene sequences with ρ expressed as the rate of gene conversion, $\rho = 2N_e r/2$. Data for individual loci are provided in Table 3.1.

	Length	π	Sites	Φ_w	I_a^s	ρ	θ_w	ρ/θ_w
Interspecies	3923	0.0568	808	2.6×10^{-15}	-	0.0097	0.0454	0.21
Intraspecies	2528	0.0018	21	0.0038	0.0018	0.0552	0.0020	27.9

pratensis has a panmictic and freely recombining population structure, or if the population was recently in linkage equilibrium and has undergone an evolutionarily modern population expansion. As might be expected given the observation of linkage equilibrium, the ratio of recombination to mutation rate, ρ/θ_w, for *S. flavogriseus* phylogroup *pratensis*, is among the highest observed for any bacterial or archaeal population (Table 3.2). It is important to note that the assumption of constant population size used to calculate ρ in LDhat has yet to be validated for microbial species. In addition, the lack of linkage disequilibrium also makes the calculation of an exact value for ρ impossible. Despite these limitations and pitfalls, this method has been used widely for microorganisms and is thought to allow meaningful comparison between microbial populations (for review see (23)). In general, departure from assumptions should result in underestimate of the recombination rate and so the value of ρ/θ_w that we provide should be treated as a lower bound. It was not possible to

calculate the number of nucleotide substitutions due to recombination or mutation (r/m) with ClonalFrame (18) due to the low levels of polymorphism and the high rates of recombination observed for *S. flavogriseus* phylogroup *pratensis*. As a result, an estimate of r/m was made using the single locus variant approach (r/m = 23.5 for sites), but this method is known to underestimate recombination rates and so this estimate should also be treated as a lower bound (24).

Interspecies homologous recombination among Streptomyces species

In addition to measuring intraspecies gene exchange, we also examined gene exchange between *Streptomyces* species using an existing MLSA dataset comprised of 53 different species affiliated with the *S. griseus* clade (15). While over-classification of species has been a problem for *Streptomyces* taxonomy (25), the housekeeping genes in this dataset had an average nucleotide identity of 90.5% (15), well below the 95% threshold that is thought to correspond with the conventional microbial species definition (26). The pairwise homeoplasy index test (Φ_w test) rejected the hypothesis of no recombination (Table 3.2), and the phylogenies for the *recA*, *atpD*, *rpoB*, *gyrB*, *trpB,* and 16S rRNA genes (Figure 3.2) were incongruent as determined by the Shimodaira–Hasegawa test (19) (Table 3.3). The ratio of recombination to mutation rate (Table 3.2) and the ratio of nucleotide substitutions due to recombination or mutation (r/m = 19.5, as determined for sites by ClonalFrame) for the concatenated data set exceed the values reported for recombination *within* most bacterial species (7), and exceed by several orders of magnitude interspecies recombination rates determined for other groups of bacteria and archaea (27, 28). Vos & Didelot (7) describe r/m values calculated with ClonalFrame to be very high for a species when they exceed 10 and report an r/m of 13.6 for *H. pylori.* Due to the number of polymorphisms in the sequences and the resulting lack of single locus variants it was

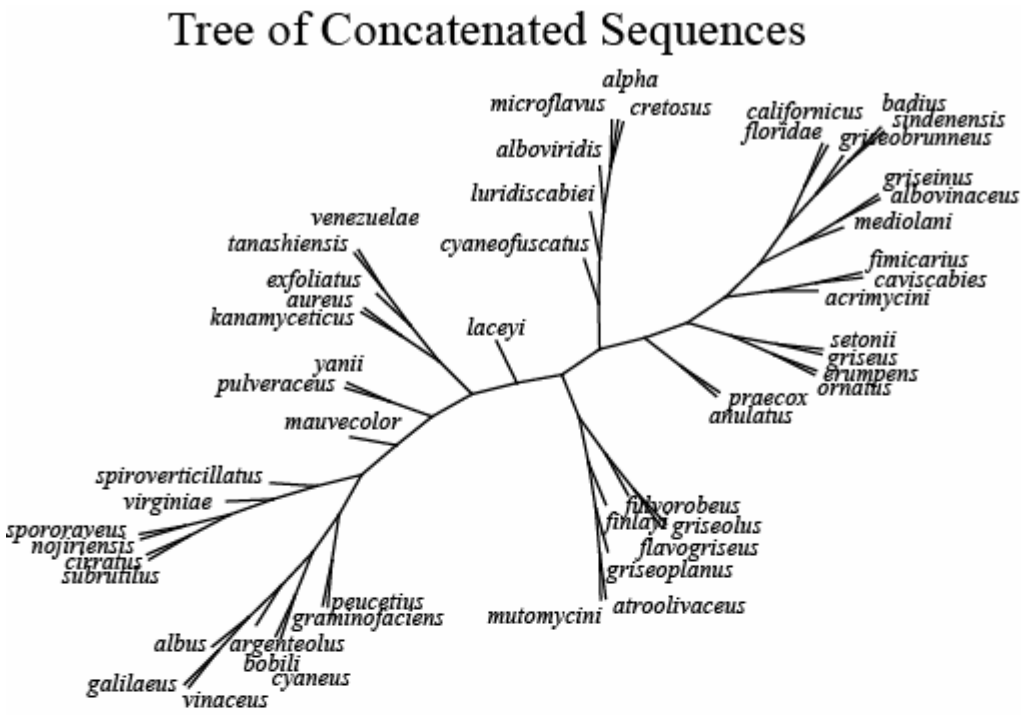
not possible to estimate r/m using the SLV method

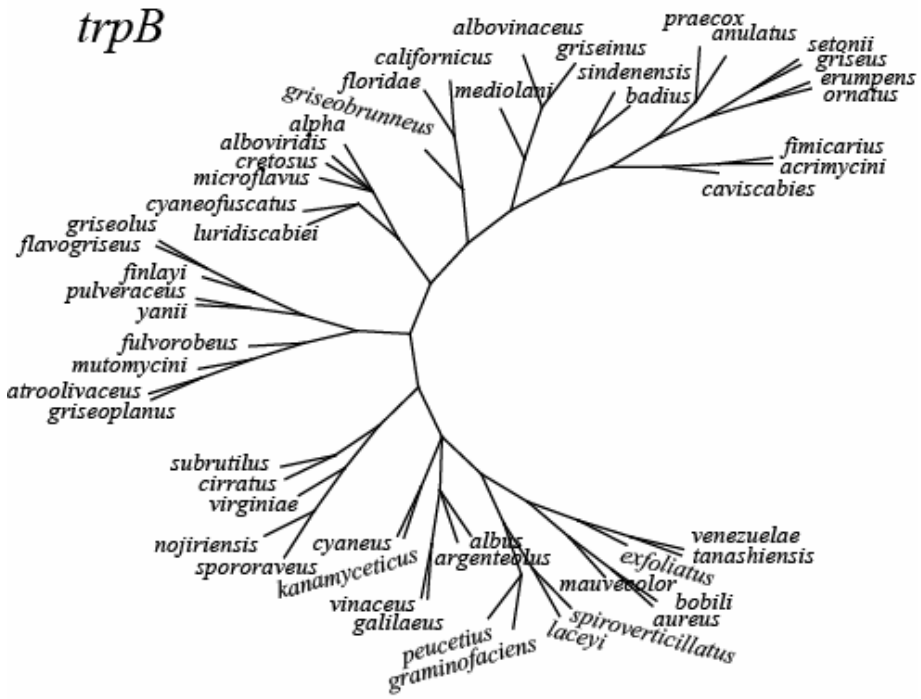
Following from these observations we used the Recombination Detection Program (RDP 2.0 (20)) to document specific instances of interspecies gene exchange. A total of 13 interspecies gene transfer events were detected, impacting 40% of the lineages examined (Figure 3.4, Table 3.4, Figure 3.3). The criteria used to detect these events are insensitive to exchange of short or similar sequences, and should be viewed as a conservative lower bound on the actual number of recombination events among these taxa. ClonalFrame (18) was used to explore the vertical pattern of inheritance in these sequences and the recombination events were mapped onto the resulting tree to contrast the vertical and horizontal patterns of inheritance (Figure 3.4). The widespread occurrence of horizontal gene exchange suggests that it is difficult to accurately depict the evolutionary history of the *Streptomyces* using vertical models of inheritance. NeighborNet analysis (29) is able to depict phylogenetic signals resulting from reticulate evolutionary processes and so this approach was also used to evaluate relationships among these taxa (Figure 3.5).

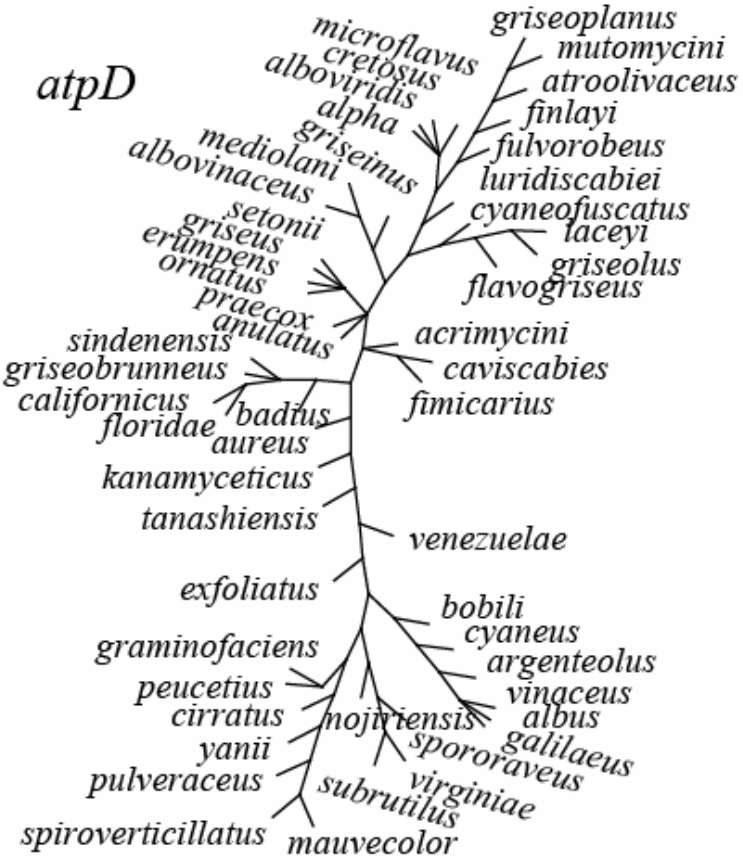
Given the extent of recombination, we chose to use Structure (22) to further evaluate the genetic structure among these named species (Figure 3.5). Structure should be sensitive to admixture among the taxa that would not be detected in the prior analysis. We present a Structure model with 3 ancestral populations, though models with 3 to 5 populations had similar likelihood and probability of the data.

It should be noted that it is difficult for Structure to infer the true number of ancestral populations for a given data set, as has been discussed by its creators (22). It is likely that more than 3 ancestral populations are contributing to the ancestry of these taxa

Figure 3.2. Maximum likelihood trees for all loci from the 53 *Streptomyces* species described in Guo et al. (15). Maximum likelihood trees were created using the tree bisection reconnection algorithm in PAUP and incongruence between trees was then evaluated using the Shimodaira-Hasegawa test (Table 3.3). In Guo *et al.*, 2008 '*griseoplanus*' was misidentified and should be '*flavogriseus* 45-CD' while '*flavogriseus*' is '*flavogriseus* Heim'.


Tree of Concatenated Sequences
alpha
microflavus
cretosus
californicus
badius
sindenensis
floridae
griseobrunneus
alboviridis
griseinus
albovinaceus
luridiscabiei
mediolani
venezuelae
tanashiensis
cyaneofuscatus
fimicarius
caviscabies
exfoliatus
aureus
acrimycini
kanamyceticus
laceyi
setonii
griseus
yanii
erumpens
pulveraceus
ornatus
praecox
anulatus
mauvecolor
spiroverticillatus
virginiae
fulvorobeus
finlayi
griseolus
sporoaveus
nojiriensis
flavogriseus
griseoplanus
cirratus
subrutilus
peucetius
atroolivaceus
graminofaciens
mutomycini
albus
argenteolus
bobili
galilaeus
cyaneus
vinaceus
16S rRNA
ornatus
fulvorobeus
erumpens
microflavus
griseus
luridiscabiei
alboviridis
subrutilus
cretosus
sporoaveus
cyaneofuscatus
alpha
cirratus
nojiriensis
flavogriseus
virginiae
griseolus
vinaceus
argenteolus
mauvecolor
albus
galilaeus
griseobrunneus
anulatus
cyaneus
praecox
bobili
setonii
acrimycini
caviscabies
spiroverticillatus
fimicarius
griseoplanus
sindenensis
albovinaceus
badius
mediolani
mutomycini
griseinus
californicus
atroolivaceus
finlayi
floridae
pulveraceus
laceyi
yanii
peucetius
graminofaciens
tanashiensis
kanamyceticus
aureus
exfoliatus
venezuelae


trpB
albovinaceus
praecox
anulatus
setonii
griseus
erumpens
ornatus
californicus
griseinus
sindenensis
floridae
mediolani
badius
griseobrunneus
alpha
alboviridis
cretosus
microflavus
fimicarius
acrimycini
caviscabies
cyaneofuscatus
griseolus
luridiscabiei
flavogriseus
finlayi
pulveraceus
yanii
fulvorobeus
mutomycini
atroolivaceus
griseoplanus
subrutilus
cirratus
virginiae
nojiriensis
cyaneus
albus
venezuelae
tanashiensis
sporoaveus
kanamyceticus
argenteolus
exfoliatus
mauvecolor
bobili
vinaceus
galilaeus
aureus
spiroverticillatus
peucetius
graminofaciens
laceyi


atpD
griseoplanus
microflavus
mutomycini
cretosus
alboviridis
atroolivaceus
alpha
finlayi
griseinus
mediolani
fulvorobeus
albovinaceus
luridiscabiei
setonii
cyaneofuscatus
griseus
laceyi
erumpens
griseolus
ornatus
flavogriseus
praecox
anulatus
acrimycini
sindenensis
caviscabies
griseobrunneus
californicus
fimicarius
badius
floridae
aureus
kanamyceticus
tanashiensis
venezuelae
exfoliatus
bobili
cyaneus
graminofaciens
argenteolus
peucetius
vinaceus
cirratus
nojiriensis
albus
galilaeus
yanii
sporoaveus
pulveraceus
virginiae
subrutilus
spiroverticillatus
mauvecolor

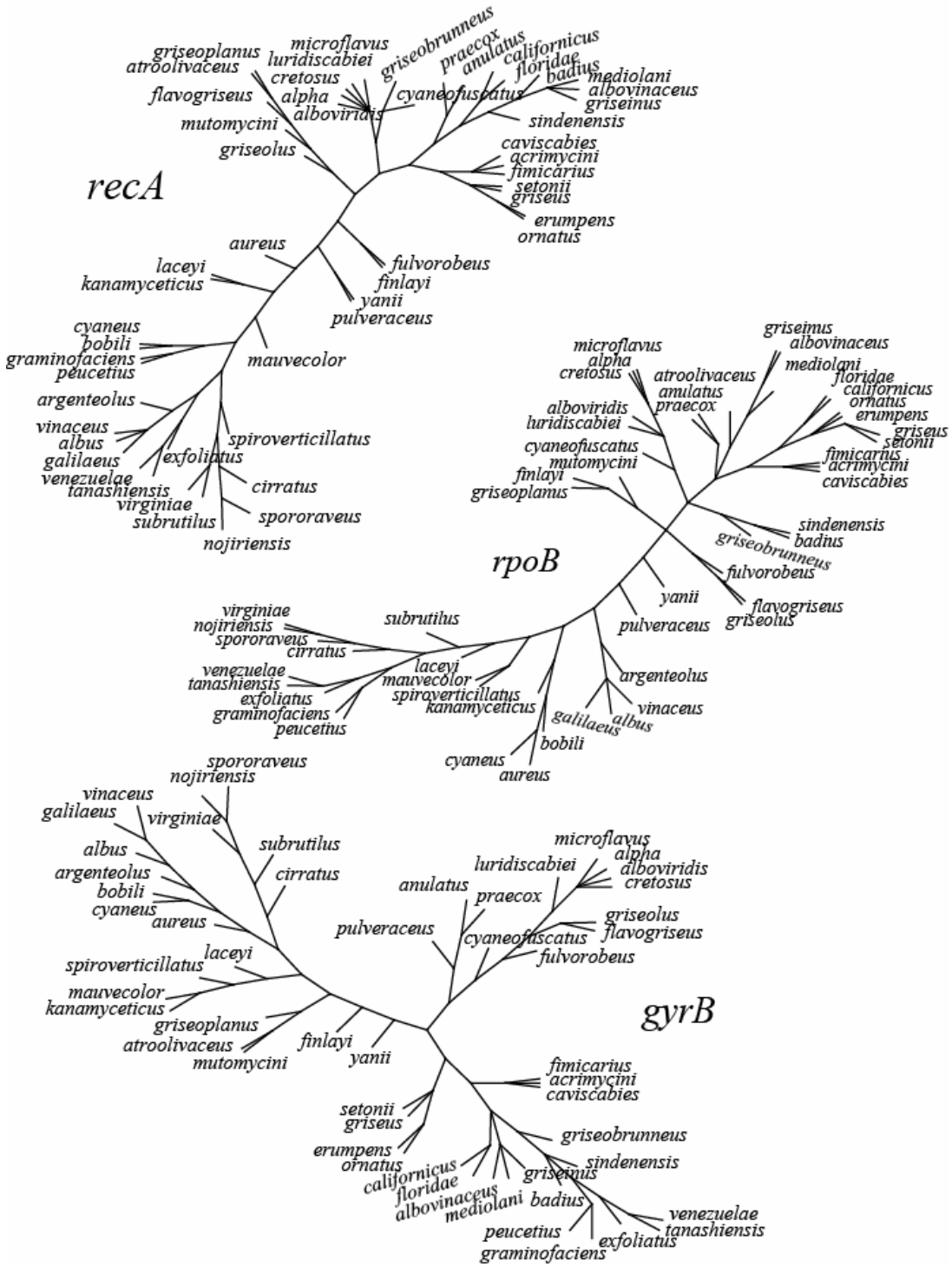

recA
griseoplanus
atroolivaceus
microflavus
luridiscabiei
cretosus
flavogriseus
alpha
alboviridis
griseobrunneus
praecox
anulatus
californicus
floridae
badius
mediolani
albovinaceus
griseinus
sindenensis
cyaneofuscatus
mutomycini
griseolus
caviscabies
acrimycini
fimicarius
setonii
griseus
erumpens
ornatus
aureus
laceyi
kanamyceticus
fulvorobeus
finlayi
yanii
pulveraceus
cyaneus
bobili
graminofaciens
peucetius
mauvecolor
argenteolus
vinaceus
albus
galilaeus
venezuelae
tanashiensis
virginiae
subrutilus
exfoliatus
spiroverticillatus
cirratus
sporoveraeus
nojiriensis
rpoB
griseinus
albovinaceus
microflavus
alpha
cretosus
atroolivaceus
amulatus
praecox
mediolani
floridae
californicus
ornatus
erumpens
griseus
setonii
alboviridis
luridiscabiei
cyaneofuscatus
mutomycini
finlayi
griseoplanus
fimicarius
acrimycini
caviscabies
sindenensis
badius
griseobrunneus
fulvorobeus
yanii
flavogriseus
griseolus
pulveraceus
virginiae
nojiriensis
sporoveraeus
cirratus
subrutilus
venezuelae
tanashiensis
exfoliatus
graminofaciens
peucetius
laceyi
mauvecolor
spiroverticillatus
kanamyceticus
argenteolus
vinaceus
galilaeus
albus
bobili
cyaneus
aureus
gyrB
sporoveraeus
nojiriensis
vinaceus
galilaeus
virginiae
albus
argenteolus
bobili
cyaneus
aureus
subrutilus
cirratus
laceyi
spiroverticillatus
mauvecolor
kanamyceticus
griseoplanus
atroolivaceus
mutomycini
finlayi
yanii
microflavus
alpha
luridiscabiei
alboviridis
cretosus
anulatus
praecox
griseolus
flavogriseus
pulveraceus
cyaneofuscatus
fulvorobeus
fimicarius
acrimycini
caviscabies
setonii
griseus
griseobrunneus
erumpens
ornatus
sindenensis
californicus
griseinus
floridae
albovinaceus
mediolani
badius
venezuelae
tanashiensis
peucetius
exfoliatus
graminofaciens

Table 3.3. Shimodaira-Hasegawa test implemented on trees from Figure 3.2. Values are the probability that the level of incongruence observed could be due to chance. Rows present the sequences used to assess tree topology, columns the tree topology being tested. 'Concat' stands for the concatenated sequence. Recombination is expected to yield significant incongruence between trees.

	Concat.	16S	*recA*	*atpD*	*rpoB*	*gyrB*	*trpB*
Concat.	--	0.000	0.000	0.000	0.000	0.000	0.000
16S	0.000	--	0.000	0.000	0.000	0.000	0.000
recA	0.020	0.000	--	0.000	0.000	0.000	0.020
atpD	0.140	0.000	0.000	--	0.000	0.000	0.000
rpoB	0.002	0.000	0.000	0.000	--	0.000	0.000
gyrB	0.001	0.000	0.000	0.000	0.000	--	0.000
trpB	0.030	0.000	0.009	0.000	0.000	0.0000	--

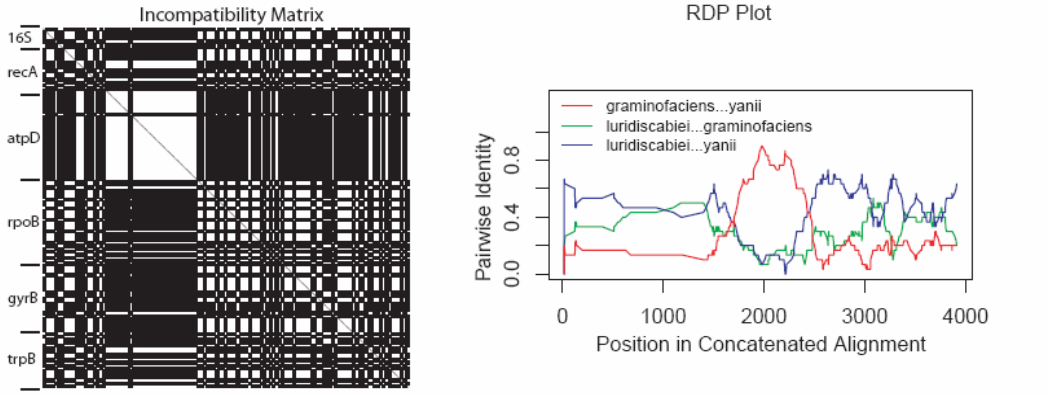


Figure 3.3. Example of graphical output from RDP 2.0 (20) and Reticulate (30) used to assess evidence for recombination, shown for the first row of Table 3.4. The incompatibility matrix presents alignments of concatenated sequences along the horizontal and vertical axes with incompatible sites indicated by black blocks in the matrix. The large block corresponding to the *atpD* gene shows clear evidence of recombination between these species. This event is also clearly evident in the three way comparison of sequence similarity performed by RDP. These figures can be used to qualitatively observe the events detected in Table 3.4.

Figure 3.4. Evidence for widespread interspecies homologous recombination among *Streptomyces* species. Recombination events supported by multiple statistical tests (Table 3.4) are mapped onto a 95% consensus ClonalFrame tree. Arrows represent recombination events, with different colors used to identify the gene or genes that were exchanged according to the legend. The direction and placement of arrows can be inferred from the events detected by RDP (Table 3.4). The 16S rRNA gene transfer event originating outside of the tree is inferred to be from an unknown donor (Table 3.4). Species names are colored to reflect hypothetical ancestral populations assigned by Structure output in Figure 3.5.

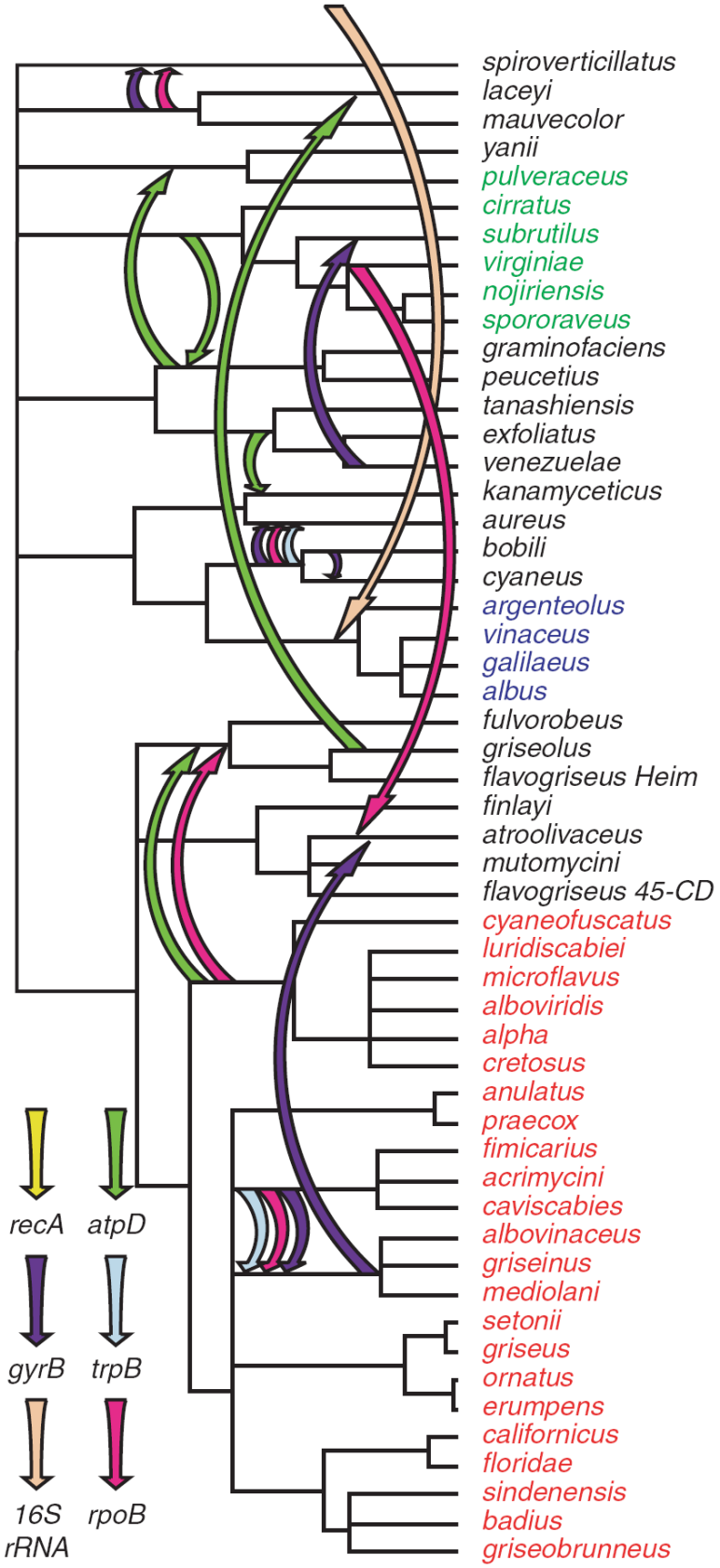

spiroverticillatus
laceyi
mauvecolor
yanii
pulveraceus
cirratus
subrutilus
virginiae
nojiriensis
sporoaveus
graminofaciens
peucetius
tanashiensis
exfoliatus
venezuelae
kanamyceticus
aureus
bobili
cyaneus
argenteolus
vinaceus
galilaeus
albus
fulvorobeus
griseolus
flavogriseus Heim
finlayi
atroolivaceus
mutomycini
flavogriseus 45-CD
cyaneofuscatus
luridiscabiei
microflavus
alboviridis
alpha
cretosus
anulatus
praecox
fimicarius
acrimycini
caviscabies
albovinaceus
griseinus
mediolani
setonii
griseus
ornatus
erumpens
californicus
floridae
sindenensis
badius
griseobrunneus
recA
atpD
gyrB
trpB
16S rRNA
rpoB

Table 3.4. Recombination events detected with RDP 2.0 (20). Columns identify the gene daughter (recipient) and donor, the sequence region, and *p*-values from 8 different statistical tests for recombination (31-37). Gene boundaries in the concatenated sequence are as follows: 16S rRNA, 1:1389; recA, 1390:1893; *atpD*, 1894:2389; *rpoB*, 2390:2929; *gyrB*, 2930:3352; *trpB*, 3353:3923. The abbreviation, u(*x*), indicates an unknown donor, with support for the existence of an unknown donor calculated through the use of sequence from isolate *x*. Heim is used for *flavogrisues* Heim, *spiroverticillatus* is abbreviated as *spirover*. and *graminofaciens* as *gramin*.

Daughter	Donor	Region	RDP	Geneconv	Bootscan	Max Chi	Chimaera	SiScan	3Seq	LARD
yanii	*gramin.*	1728:2391	$6.2x10^{-10}$	$1.1x10^{-6}$	$1.4x10^{-9}$	$5.2x10^{-5}$	$1.3x10^{-6}$	$2.1x10^{-6}$	$5.2x10^{-10}$	$3.2x10^{-30}$
pulveraceus	*gramin.*	1728:2391	$6.2x10^{-10}$	$1.1x10^{-6}$	$1.4x10^{-9}$	$5.2x10^{-5}$	$1.3x10^{-6}$	$2.1x10^{-6}$	$5.2x10^{-10}$	$3.2x10^{-30}$
galilaeus	u(Heim)	151:1449	$1.2x10^{-6}$	$2.5x10^{-11}$	$4.6x10^{-10}$	$1.0x10^{-6}$	$1.0x10^{-6}$	$2.8x10^{-7}$	$5.8x10^{-12}$	$4.1x10^{-39}$
albus	u(Heim)	151:1449	$1.2x10^{-6}$	$2.5x10^{-11}$	$4.6x10^{-10}$	$1.0x10^{-6}$	$1.0x10^{-6}$	$2.8x10^{-7}$	$5.8x10^{-12}$	$4.1x10^{-39}$
vinaceus	u(Heim)	151:1449	$1.2x10^{-6}$	$2.5x10^{-11}$	$4.6x10^{-10}$	$1.0x10^{-6}$	$1.0x10^{-6}$	$2.8x10^{-7}$	$5.8x10^{-12}$	$4.1x10^{-39}$
argenteolus	u(Heim)	151:1449	$1.2x10^{-6}$	$2.5x10^{-11}$	$4.6x10^{-10}$	$1.0x10^{-6}$	$1.0x10^{-6}$	$2.8x10^{-7}$	$5.8x10^{-12}$	$4.1x10^{-39}$
spiroverticill.	u(Heim)	151:1449	$1.2x10^{-6}$	$2.5x10^{-11}$	$4.6x10^{-10}$	$1.0x10^{-6}$	$1.0x10^{-6}$	$2.8x10^{-7}$	$5.8x10^{-12}$	$4.1x10^{-39}$
aureus	*bobili*	2484:3807	$2.5x10^{-5}$	NS	NS	$3.6x10^{-8}$	$1.5x10^{-9}$	$1.7x10^{-3}$	$1.9x10^{-9}$	$7.1x10^{-36}$
gramin.	*nojiriensis*	1893:2300	$4.0x10^{-7}$	$6.7x10^{-7}$	$1.9x10^{-8}$	$6.9x10^{-4}$	$3.7x10^{-3}$	$4.9x10^{-7}$	$2.1x10^{-2}$	$1.0x10^{-35}$
peucetius	*nojiriensis*	1893:2300	$4.0x10^{-7}$	$6.7x10^{-7}$	$1.9x10^{-8}$	$6.9x10^{-4}$	$3.7x10^{-3}$	$4.9x10^{-7}$	$2.1x10^{-2}$	$1.0x10^{-35}$
atroolivaceus	*mediolani*	2920:3339	$1.6x10^{-6}$	$3.9x10^{-5}$	$8.4x10^{-5}$	$2.8x10^{-8}$	$3.5x10^{-7}$	$1.2x10^{-5}$	$1.1x10^{-6}$	$2.4x10^{-25}$
Heim	*cyaneofuscatus*	1812:2898	$1.3x10^{-6}$	$1.3x10^{-4}$	$6.5x10^{-7}$	$6.6x10^{-3}$	$2.0x10^{-4}$	$1.5x10^{-8}$	$2.2x10^{-5}$	$1.6x10^{-18}$
griseolus	*cyaneofuscatus*	1812:2898	$1.3x10^{-6}$	$1.3x10^{-4}$	$6.5x10^{-7}$	$6.6x10^{-3}$	$2.0x10^{-4}$	$1.5x10^{-8}$	$2.2x10^{-5}$	$1.6x10^{-18}$
fulvorobeus	*cyaneofuscatus*	1812:2898	$1.3x10^{-6}$	$1.3x10^{-4}$	$6.5x10^{-7}$	$6.6x10^{-3}$	$2.0x10^{-4}$	$1.5x10^{-8}$	$2.2x10^{-5}$	$1.6x10^{-18}$
laceyi	*Heim*	1757:2391	$1.3x10^{-7}$	$9.5x10^{-7}$	$1.2x10^{-9}$	$6.2x10^{-6}$	$1.1x10^{-4}$	$1.0x10^{-7}$	$5.3x10^{-3}$	$2.6x10^{-33}$
spirovert.	*laceyi*	2292:3627	NS	NS	NS	NS	$1.2x10^{-6}$	NS	NS	$3.0x10^{-30}$
subrutilus	*venezuelae*	2987:3263	NS	NS	NS	$2.3x10^{-6}$	$2.0x10^{-3}$	NS	NS	$1.3x10^{-27}$
albovinaceus	*caviscabies*	2598:3753	NS	NS	NS	$4.2x10^{-4}$	$2.5x10^{-2}$	$1.1x10^{-7}$	$2.5x10^{-2}$	$7.0x10^{-10}$
griseinus	*caviscabies*	2598:3753	NS	NS	NS	$4.2x10^{-4}$	$2.5x10^{-2}$	$1.1x10^{-7}$	$2.5x10^{-2}$	$7.0x10^{-10}$
mediolani	*caviscabies*	2598:3753	NS	NS	NS	$4.2x10^{-4}$	$2.5x10^{-2}$	$1.1x10^{-7}$	$2.5x10^{-2}$	$7.0x10^{-10}$
kanamyceticus	*venezuelae*	1892:2491	NS	NS	$6.2x10^{-2}$	$2.9x10^{-4}$	NS	$8.6x10^{-6}$	NS	$1.6x10^{-15}$
atroolivaceus	*virginiae*	2398:2820	$6.1x10^{-2}$	NS	$1.2x10^{-3}$	$1.4x10^{-2}$	NS	NS	NS	$3.9x10^{-15}$
cyaneus	*bobili*	3066:3487	NS	NS	NS	$4.8x10^{-3}$	$1.6x10^{-3}$	$3.7x10^{-8}$	NS	$5.6x10^{-18}$

and that further genetic structure will be revealed by analysis of a greater diversity of *Streptomyces* and a greater diversity of loci. The purpose of this analysis was to examine patterns of admixture resulting from horizontal gene exchange and $k = 3$ was selected because it represents the smallest value for the number of ancestral

populations that captured meaningful structure in the data and was supported by the likelihood and probability of the data (as discussed (22)). The pattern of ancestry generated by the Structure model was generally consistent with the recombination events depicted in Figure 3.4, while providing additional evidence for admixture in a range of taxa. Two of the ancestral populations in the Structure model correspond to species clusters that have been proposed previously based on phenotypic properties (38) (see Figure 3.5 legend for details). In several cases, lineages inferred to be admixed correspond to species for which various genetic and phenotypic markers have provided incongruent phylogenetic and taxonomic information (25, 38). For example, the species *S. atroolivaceus*, *S. finlayi*, *S. flavogriseus*, and *S. griseolus,* have been placed in as many as 4 separate species clusters on the basis of phenotypic characteristics (38). Our results suggest that these 4 taxa contain housekeeping genes whose ancestry originates in the *S. griseus* and *S. lavendulae* clades, suggesting that reticulate processes have played a major role in the evolution of these species.

The three ancestral populations found by Structure were also recapitulated in the NeighborNet analysis (Figure 3.5), with species inferred to be admixed by Structure found to occupy positions intermediate to the three ancestral populations in the NeighborNet network. Reticulation in the phylogenetic network generally supports the findings from Structure but cannot provide confirmation since reticulation in such networks may result either from genuine gene exchange or may be due to uncertainty in the phylogenetic signal captured in the data. Thus, it is possible that some of the lineages in Figure 3.5 are not genuinely admixed but are inferred to be admixed due to a failure to find true structure in the data. We evaluated Structure models with 3 to 10 ancestral populations (data not shown) and found that the degree of admixture tended

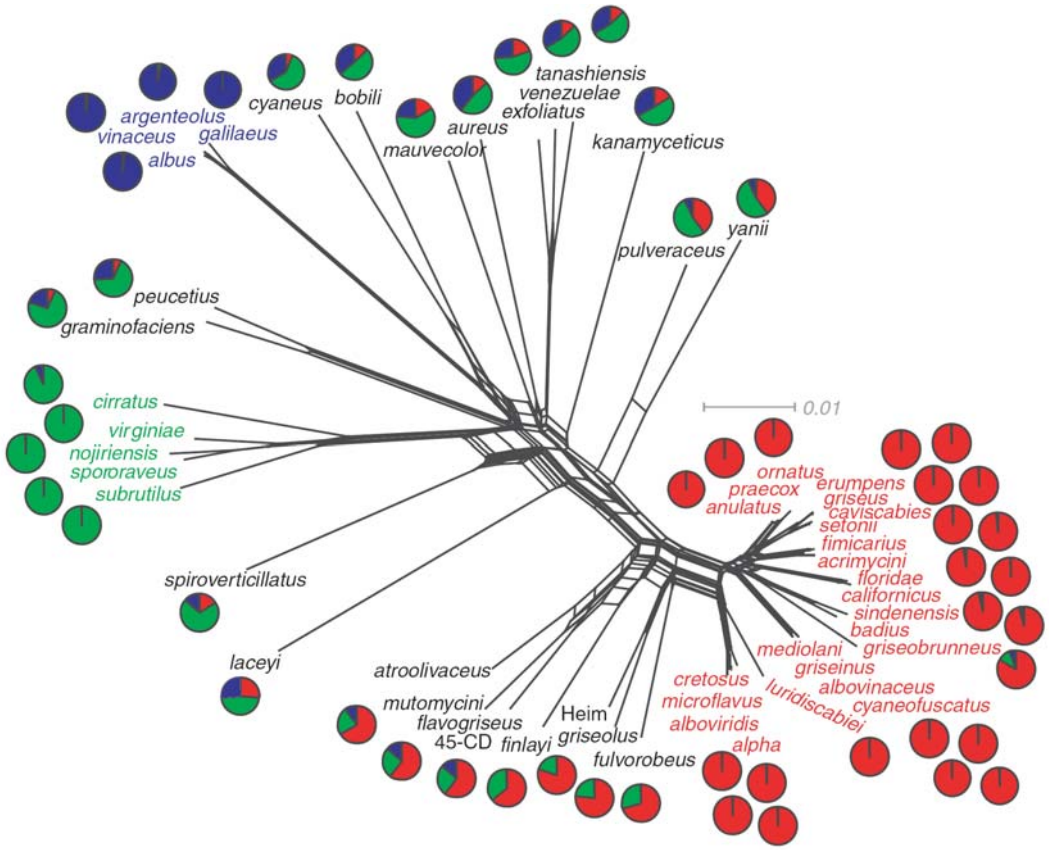


Figure 3.5. NeighborNet and Structure analyses of Guo *et al*. (15) MLSA data. Boxes in the network represent uncertainty in the phylogeny and are expected if horizontal gene exchange has occurred. The scale bar is equal to 1% sequence divergence. The output from Structure analysis is indicated in the pie charts with colors indicating the proportion of ancestry estimated from each of three hypothetical ancestral populations. Two of the ancestral populations map onto species clusters that have been described previously (38): the *S. griseus* clade (red) and the S. *lavendulae* clade (green). Heim is used for *S. flavogriseus* Heim.

to increase asymmetrically across lineages with the number of ancestral populations included in the model, suggesting that evidence for admixture obtained with $k = 3$ is not an artifact of having too few populations in the model. Admixture remained largely absent from species in the *S. griseus* cluster regardless of the Structure model. In addition, the species: *atroolivaceus, mutomycini, finlayi, flavogriseus, griseolus,* and *fulvorobeus* were always composed of a mixture of two ancestral populations with a dominant contribution from the *S. griseus* ancestral population. These species clearly show asymmetrical distribution of ancestral populations (regardless of the structure model used) with the greatest contribution from the most closely related lineage found in both the ClonalFrame tree (Figure 3.4) and the NeighborNet analysis (Figure 3.5).

Implications

The influence of horizontal gene exchange on the evolution of the *Streptomyces* appears to be profound. These data suggest that reticulation is widespread in the *Streptomyces* phylogeny, with new lineages potentially arising as a result of hybridization between species clusters. There are reasons to suggest that the dsDNA conjugation system of streptomycetes may be associated with high levels of gene exchange among microbial populations. This conjugation system permits interspecies recombination between isolates in the laboratory (39), and can generate hybrid strains with genomes having nearly equal genetic contributions from each parent (40). Such laboratory generated hybrids are described to display new combinations of parent phenotypes including changes in phage sensitivity (41) and antibiotic production (42). These observations may go a long way toward explaining why the taxonomy of the streptomycetes has been difficult to resolve historically. Variation in morphological, physiological, and biochemical characteristics both within and between named species of the *Streptomyces* cause incongruence between phenotypic and genotypic groupings

(25). Given the present data it seems fair to hypothesize that this incongruence is due to horizontal exchange of phylogenetic markers and genes that encode particular phenotypic traits. Reticulate evolutionary processes are likely to blur the boundaries between *Streptomyces* species with gene exchange producing metaspecies that are difficult to classify. Fuzzy species boundaries and metaspecies have been indicated within other microbial groups (28, 43) and are well known in plants and animals (44). The hypothesis of widespread gene exchange among streptomycetes would explain existing taxonomic inconsistencies among these organisms and would provide an evolutionary framework that could facilitate an understanding of their taxomony and phylogeny.

It is interesting to note that the intraspecies recombination rate within *S. flavogriseus* phylogroup *pratensis* exceeded the interspecies recombination rate by more than two orders of magnitude (Table 3.2). These data are consistent with the idea that recombination is acting as a cohesive force which declines in strength with increasing sequence divergence (5). While that may be the case, it is important to consider that the different *Streptomyces* species examined in this study were isolated from a wide range of localities and that information on their biogeography is not available. Thus, the contrast between the intraspecies and interspecies recombination rates could be a consequence of geographical or ecological barriers to gene exchange and not simply a function of sequence divergence. The stark discontinuity that we observed between rates of interspecies and intraspecies gene exchange would be expected to promote the persistence of cohesive genetic clusters. Given the high rate of interspecies gene exchange observed, we might expect the frequent introduction of foreign homologous genes into a population, with these genes frequently eliminated either by genetic drift or the cohesive effects of recombination. Introgression of foreign genes into the

population or establishment of hybrid populations would be expected to result if the foreign genes have adaptive significance or as a consequence of demographic phenomena. To determine whether streptomycetes form cohesive genetic clusters that might be properly described as species, or rather represent points along a continuum of genetic exchange within the genus, will require investigation of multiple sympatric populations.

It is clear that further investigation of the population structure and biogeography of streptomycetes is needed to understand the ecological and evolutionary causes and consequences of these high rates of gene exchange. Such data should help to resolve longstanding inconsistencies in our understanding of streptomycete phylogeny and taxonomy. These issues are not merely of academic concern as streptomycetes are a preeminent source of antibiotics and bioactive compounds and progress in understanding the ecology and evolution of antibiotic production has been limited by our lack of a coherent phylogenetic framework for these organisms. In addition, an understanding of the biogeography of these organisms should lead to the development of rational sampling strategies for discovering novel genetic diversity and bioactive compounds within natural populations of *Streptomyces*.

REFERENCES

1. Doolittle WF & Bapteste E (2007) Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* 104(7):2043-2049.
2. Fraser C, Alm EJ, Polz MF, Spratt BG, & Hanage WP (2009) The bacterial species challenge: Making sense of genetic and ecological diversity. *Science* 323(5915):741-746.
3. Koeppel A*, et al.* (2008) Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. *Proc. Natl. Acad. Sci. U. S. A.* 105(7):2504-2509.
4. Hunt DE*, et al.* (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320(5879):1081-1085.
5. Fraser C, Hanage WP, & Spratt BG (2007) Recombination and the nature of bacterial speciation. *Science* 315:476-480.
6. Sheppard SK, McCarthy ND, Falush D, & Maiden MCJ (2008) Convergence of Campylobacter species: Implications for bacterial evolution. *Science* 320(5873):237-239.
7. Vos M & Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 3(2):199-208.
8. Hopwood DA (2006) Soil to genomics: The *Streptomyces* chromosome. *Ann. Rev. Genet.* 40:1-23.
9. Kataoka M, Seki T, & Yoshida T (1991) Regulation and function of the *Streptomyces* plasmid psn22 genes involved in pock formation and inviability. *J. Bacteriol.* 173(24):7975-7981.
10. Hopwood DA, Lydiate DJ, Malpartida F, & Wright HM (1985) Conjugative sex plasmids of Streptomyces. *Basic Life Sci.* 30:615-634.

11. Chater KF, Hopwood DA, Kieser T, & Thompson CJ (1982) Gene cloning in Streptomyces. *Curr. Top. Microbiol. Immunol.* 96:69-95.
12. Elnakeeb MA & Lecheval.Ha (1963) Selective Isolation of Aerobic Actinomycetes. *Appl. Microbiol.* 11(2):75-&.
13. Shirling EB & Gottlieb D (1970) Report of the International Streptomyces Project 5 Years of Collaborative Research. *Prauser, H. (Edited by). The Actinomycetales. The Jena International Symposium on Taxonomy. 439 p. Illus. Veb Gustav Fischer Verlag: Jena, E. Germany*:79-89.
14. Wink J (2009) The Compendium of Actinomycetales. (Sanofi-Aventis Deutschland GmbH).
15. Guo YP, Zheng W, Rong XY, & Huang Y (2008) A multilocus phylogeny of the *Streptomyces griseus* 16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. *Int. J. Syst. Evol. Microbiol.* 58:149-159.
16. Haubold B & Hudson RR (2000) LIAN 3.0: detecting linkage disequilibrium in multilocus data. *Bioinformatics* 16(9):847-848.
17. McVean G, Awadalla P, & Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160(3):1231-1241.
18. Didelot X & Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175:1251-1266.
19. Shimodaira H & Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16(8):1114-1116.
20. Martin DP, Williamson C, & Posada D (2005) RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* 21(2):260-262.
21. Huson DH & Bryant D (2006) Application of phylogenetic networks in

evolutionary studies. *Mol. Biol. Evol.* 23(2):254-267.

22. Falush D, Stephens M, & Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 164(4):1567-1587.

23. Perez-Losada M*, et al.* (2006) Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect. Genet. Evol.* 6(2):97-112.

24. Feil EJ, Maiden MCJ, Achtman M, & Spratt BG (1999) The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol. Biol. Evol.*16(11):1496-1502.

25. Anderson AS & Wellington EMH (2001) The taxonomy of *Streptomyces* and related genera. *Int. J. Syst. Evol. Microbiol.* 51:797-814.

26. Konstantinidis KT & Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102(7):2567-2572.

27. Eppley JM, Tyson GW, Getz WM, & Banfield JF (2007) Genetic exchange across a species boundary in the archaeal genus *Ferroplasma*. *Genetics* 177:407-416.

28. Papke RT*, et al.* (2007) Searching for species in haloarchaea. *Proc. Natl. Acad. Sci. U. S. A.* 104(35):14092-14097.

29. Bryant D & Moulton V (2004) Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.*21(2):255-265.

30. Jakobsen IB & Easteal S (1996) A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput. Appl. Biosci.* 12(4):291-295.

31. Boni MF, Posada D, & Feldman MW (2007) An exact nonparametric method

for inferring mosaic structure in sequence triplets. *Genetics* 176:1035-1047.

32. Martin DP, Posada D, Crandall KA, & Williamson C (2005) A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res. Hum. Retroviruses* 21(1):98-102.

33. Posada D & Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proc. Natl. Acad. Sci. U. S. A.* 98(24):13757-13762.

34. Gibbs MJ, Armstrong JS, & Gibbs AJ (2000) Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* 16(7):573-582.

35. Holmes EC, Worobey M, & Rambaut A (1999) Phylogenetic evidence for recombination in dengue virus. *Mol. Biol. Evol.*16(3):405-409.

36. Padidam M, Sawyer S, & Fauquet CM (1999) Possible emergence of new geminiviruses by frequent recombination. *Virology* 265(2):218-225.

37. Smith JM (1992) Analyzing the mosaic structure of genes. *J. Mol. Evol.* 34(2):126-129.

38. Kampfer P, Kroppenstedt RM, & Dott W (1991) A numerical classification of the genera *Streptomyces* and *Streptoverticillium* using miniaturized physiological tests. *J. Gen. Microbiol.* 137:1831-1891.

39. Alacevic M (1963) Interspecific recombination in *Streptomyces*. *Nature* 197:1323.

40. Wang SJ, Chang HM, Lin YS, Huang CH, & Chen CW (1999) *Streptomyces* genomes: circular genetic maps from the linear chromosomes. *Microbiology* 145:2209-2220.

41. Lomovskaya ND, Voeykova TA, & Mkrtumian NM (1977) Construction and properties of hybrids obtained in interspecific crosses between *Streptomyces*

coelicolor A3(2) and *Streptomyces griseus* Kr15. *J. Gen. Microbiol.* 98:187-198.

42. Stoycheva Z, Todorov T, & Peltekova V (1994) Intergeneric crosses between *Streptomyces ambofaciens* and *Saccharopolyspora erythraea*. *Folia Microbiologica* 39(1):13-18.

43. Hanage WP, Fraser C, & Spratt BG (2005) Fuzzy species among recombinogenic bacteria. *BMC Biol.* 3.

44. Mallet J (2008) Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* 363(1506):2971.

CHAPTER 4

GENETIC AND PHENOTYPIC CHARACTERIZATION OF A POPULATION OF *STREPTOMYCES FLAVOGRISEUS* PHYLOGROUP *PRATENSIS*

Introduction

Streptomyces produce the majority of clinically useful antibiotics and many other industrially relevant natural products. Antibiotic resistance is a menacing problem and a post-antibiotic era is not a fringe concern (1). However, many major pharmaceutical companies have ceased their natural products discovery work (2). Decades of screening have caused pipelines to dry up as novel compounds have become fewer and farther between without the adoption of transformative new screening methods. The alternative methods that have replaced natural product screening – such as high-throughput screening of chemical libraries against specific bacterial proteins – are failing to produce as many lead compounds as hoped (3). The small molecules produced by *Streptomyces* exist through natural selection; this resource should not yet be forgotten.

Genomic surveys do not suggest that natural product diversity has been fully sampled, but rather that laboratory screening methods fail to reveal the full biosynthetic potential of microbes. Prior to analysis of the *S. coelicolor* genome, four secondary metabolite gene clusters had been studied in this workhorse of molecular genetics. Genome analysis revealed the presence of 18 additional biosynthetic gene clusters (4). Genomes for *S. griseus* (5) and *S. avermitilis* (6) revealed 34 and 30 secondary metabolite gene clusters, respectively. It is possible that the next wave of natural

product discovery will be driven by genomic data. This is not unprecedented: classical genetics started with phenotype identification and proceeded to identifying genes, and so called reverse genetics starts with a gene and proceeds to identify its effect on phenotype.

Delineating gene pools is essential for genomic driven natural products discovery to proceed efficiently. Because genetic variation within a species is the rule rather than the exception (e.g. (7-9)), using phenotypic traits as a guide for genome mining will result in wasted time and money, and possibly the under exploration of groups more diverse than one would currently predict. A genetic survey of all named species would provide a good foundation for comparison and further exploration. Population surveys are also essential to reveal the extent of diversity within any given group and whether or not species boundaries exist.

Currently our knowledge of *Streptomyces* diversity is lacking a solid genetic framework. In the late 1960's, the International *Streptomyces* Project undertook the phenotypic classification of 450 streptomycetes and other strains that might have belonged to the genus, using nine carbon-source utilization tests, spore morphology, colony coloration and any additional pigmentation details (10). These results are still a useful source of information when examining unknown streptomycetes, especially with the use of a key (11). Two more landmark studies, Williams *et al* (12) and Kämpfer *et al* (13), used phenotypic tests to classify 475 and 821 strains of *Streptomyces* and representatives of related genera, respectively. Williams *et al*. found evidence for 19 major and 40 minor clusters, while Kämpfer found 15 major and 34 minor clusters. Both studies also found a number of single member clusters. The recommendation from these studies is to consider these clusters as species-groups,

"until further information is available" (12). Genetic surveys of the genus using multilocus sequence analysis have made clear arguments for removing the taxonomic status of several species given identical or nearly identical sequences at six housekeeping genes (14-16). Using these same data sets we previously reported high rates of interspecies recombination and a field survey revealed extraordinarily high rates of homologous recombination within one population (17). This population of *Streptomyces flavogriseus* phylogroup *pratensis* and the type strain of this group, *Streptomyces flavogriseus* IAF 45-CD (18), are the subject of this study.

Lateral gene transfer can take two forms within bacteria. There is homologous recombination, mentioned above, that involves a stretch of the donor's DNA that replaces a homologous stretch of the recipient's genome. Illegitimate, or nonhomologous, recombination (NR) is the second form, involving incorporation of a novel stretch of DNA into the recipient's genome. NR can have an enormous impact on genetic diversity within a species. In 61 sequenced *E. coli* genomes, for example, only 20% of genes in any given isolate are found in all genomes, meaning 80% are variable among strains (8). The fact that most species' pan genomes are much larger than their core genomes, and that lateral gene transfer is responsible for the variety of genes an individual may contain, led us to hypothesize that a recombinant population can exhibit a large degree of phenotypic variation, but the level of variation will depend upon the trait.

The 37 strains of *Streptomyces flavogriseus* phylogroup *pratensis* were isolated from across New York and Michigan. The type strain was isolated in Laval, Quebec, and was first reported in 1980 (18). For this study we have examined phenotypic and genetic traits of these 38 isolates. Phenotypic tests were performed for carbon-source

utilization, antibiotic resistance, and antibiosis against three test strains. Using the draft genome of *S. flavogriseus* strain IAF 45-CD, 9 PKS or NRPS clusters were discovered in addition to *traB*, the only gene whose product is essential for plasmid transfer during mating. The presence or absence of these 10 genetic traits was determined across all 38 strains. We find that for this population there is a consensus, or common to the majority, set of traits, but that most isolates deviate from the norm in at least one trait. The level of variation does depend upon the trait being examined, and phenotypic variation involving antibiotic resistance or production is higher than for other measured traits.

Materials and Methods

Isolation source

Isolates originate from: Caldwell Field, Ithaca, NY (#1,3), Chazy, NY (#2), Harford, NY (#9-11), Mitchell Street, Ithaca, NY, (#12-18), and Willsboro, NY (#19-37). Please see Chapter 3 for more details on isolation.

Carbon source tests

Carbon source utilization was tested by checking for growth on International *Streptomyces* Project medium 9 (19). Agar for these tests was washed four times using reverse osmosis purified H_2O. Screens were performed in triplicate in 96 well flat bottom plates. Isolates were grown in 2mL of tryptone-yeast extract broth (ISP medium 1) for 3 days and washed by pelleting, resuspending in 2mL PBS, and pelleting before a final resuspension in 1mL H_2O. In the three wells of the same substrate for each isolate, 10, 20 and 30ul of washed cells was used as inoculum. Carbon sources used were: α-D-glucose, L-arabinose, sucrose, D-xylose, *myo*-inositol, D-mannitol, D-fructose, rhamnose, raffinose and cellobiose.

Antibiotic resistance assays

Penicillin (P, 10 units), kanamycin (K, 30ug), chloramphenicol (C, 30ug), tetracycline (TE, 30ug), ampicillin (A, 10ug), streptomycin (ST, 10ug), erythromycin (E, 15ug) and oxytetracycline (T, 30ug) impregnated discs (Oxoid) were used to test for antibiotic resistance. In the first screen, a loop of spores was streaked across a 100mm glycerol arginine plate for each isolate, and all antibiotic discs except for chloramphenicol were spaced evenly across the plate. This provided clear results for K, AM, S, E, and T, but proved ambiguous C, TE, and P. To decrease variability between tests, spore preparations for each isolate were used to inoculate three 60mm glycerol arginine plate with 10^6 spores and one disc of C, TE or P was placed on each plate. Resistance was assessed after 7 days.

Antibiosis

Isolates were streaked onto half of a glycerol arginine plate at 30°C until sporulation was initiated, usually seven to ten days. A 25mL overlay of 0.75% LB agar was then added and after solidifying overnight at room temperature, the same isolate was streaked on the half of the plate directly over the initial streak. After sporulation of the streak on LB, the three test isolates, *Escherichia coli*, *Bacillus subtilis*, and *Mycobacterium smegmatis*, were streaked perpendicular to the *pratensis* isolate and grown at 37°C overnight. Inhibition of *B. subtilis* and *E. coli* were recorded and the plates were incubated for a further 24 hours, at which time inhibition of *M. smegmatis* was determined. Test was modified from Nkanga and Hagedorn (20).

PKS II screen using degenerate primers

Prior to the draft genome of *Streptomyces flavogriseus* coming to our attention, the

presence and sequence of polyketide synthase type II genes was determined using the degenerate primers 540F 5'-GGITGCACSTCIGGIMTSGAC-3' and 1100R 5'-CCGATSGCICCSAGIGAGTG-3' (21). Each 25 µl PCR reaction contained 10.6 µl of H2O, 2.5 µl 10x AmpliTaq Gold Buffer, 2.5 µl $MgCl_2$, 2 µl solution with 2.5 mM each dNTP, 2.5 µl 10 µM forward primer, 2.5 µl 10 µM reverse primer, 0.4 µl BSA, 1 µl of 5U/µl Taq. The product of this reaction will be referred to as pks1 (polyketide synthase #1) throughout. The PCR product was cleaned using EXOSAP, and sequencing was performed at the Cornell Life Sciences Core Laboratories Center.

traB and PKS/NRPS screen

The publicly available draft genome sequence of *Streptomyces flavogriseus* ATCC 33331 was used as input for the program NP.searcher (22). NP.searcher found five modular NRPS gene clusters, one modular PKS cluster, and one mixed modular NRPS/PKS gene cluster. A BLAST search revealed the presence of an additional PKS type II gene. Throughout this publication biosynthetic gene clusters will be labeled as np1-np7 in the order of NP.searcher output, and pks2 is the other PKS type II found through BLAST. To test the abundance of np1-np7 and pks2 in the population, seven primer sets were designed to specifically target these gene clusters. Most primer sets were intentionally designed to span domains or genes, to prevent a possible false positive if resorting of the same genes occurred to produce novel products. Table 4.1 contains a list of the primers, the natural product they are intended to hit, and the domains spanned by the primers. Primer specificity was checked using BLAST against the *S. flavogriseus* BLAST database. Each 25 µl PCR reaction contained 11.75 µl of H2O, 2.5 µl 10x AmpliTaq Gold Buffer, 3.0 µl $MgCl_2$, 2.0 µl solution with 2.5 mM each dNTP, 1.0 µl 10 µM forward primer, 1.0 µl 10 µM reverse primer, 2.5 µl DMSO, 0.25 µl of 5U/µl AmpliTaq, and 1 µl template.

The *traB* sequence from *S. flavogriseus* was found using PSI-BLAST with *traB* from *S. coelicolor* plasmid SCP1 as a starting query (23). Primers used for *traB* are forward 5'-TCCAGCTGAAGCTGAAGAAA-3' and reverse 5'-TCTGGTGGAGCT-TGCTGAC-3'.

Outgroup comparisons

Type strains from other species with closely related 16S rRNA sequences were found using BLAST searches against the nr database (24). Species and sequence accession numbers used in comparison: *S. griseus* subsp. griseus, AB030567.1; *S. luridiscabiei*, NR_025155.1; *S. caviscabies*, AF112160.1; *S. flavogriseus*, AJ494864.1; *S. ornatus*, X79326.1; *S. globisporus*, EF178686.1; *S. anulatus*, DQ026637.1; *S. fimicarius*, AY999784.1; *S. baarnensis*, EF178688.1; *S. cavourensis* subsp. washingtonensis, DQ026671.1; *S. badius*, AY999783.1; and *S. microflavus*, DQ445795.1. Sequences were aligned with ClustalW (25) and manually checked for quality. DNA distance matrices were made using Phylip's DNADist program (26).

Table 4.1. Primers used to survey biosynthetic gene cluster occurrence. Table columns are for assigned numbers for PKS and NRPS containing gene clusters, domains spanned by the PCR product, and the forward and reverse primers.

NP	Domains	Forward primer	Reverse primer
1	A	GTCACCCAGCAGTTCTCCAT	CGGTGTCCACTGATCTTGAC
2	A-PCP-C	CCAACACACGTGCCTACATC	TGAAGTGGAAACTGCTGACG
3	KS-AT	GGCTCCATCAAGTCCAACAT	AGTACTCCGGGTCGGTGAG
4	AT-DH	CTTCACCGACTGGGACCTC	GGTGAGTTCGGTGACGTGT
5	A-PCP-C	GACGGACAGGTGAAACTGC	AGGGTGTAGTCGGCGTACTG
6	A-PCP-C	CAGGTCTCACTCGGCTACCT	AGCAGAATCAGCCAGGACAC
7	A-PCP-C	TCTACGGACAGACGGAGACC	GAGCGACAGGGAGGTTGTAG
pks2	KS-KS	GACTGGTAGGCGCTGACGTA	CACGTCTACGCCGAGATAGG

Results

Carbon source tests

Results of carbon-source growth utilization assays are in Table 4.2. Eight isolates vary in the ability to utilize one carbon-source each and none vary from the majority at two or more carbon-sources. Fructose is the carbon-source that is most variably utilized, with 7.9% variation. On average, each carbon source will be utilized differentially by 2.1% of the population. The type strain and majority carbon-source utilization profile is: (+) glucose, (+) xylose, (+) arabinose, (−) sucrose, (−) raffinose, (+) rhamnose, (+) mannitol, (−) inositol, (+) fructose, (+) cellobiose.

Antibiotic resistance tests

Antibiotic resistance (Table 4.3) showed greater variation than did carbon-source utilization. Tetracycline was the most variable, with 29% sensitive and 71% resistant. All strains were resistant to penicillin and ampicillin, and all strains were sensitive to kanamycin and streptomycin. For the variable antibiotics, the majority of isolates were resistant to tetracycline and oxytetracycline, and sensitive to chloramphenicol and erythromycin, to which only one isolate was resistant.

Antibiosis

Antibiotic production, assayed by affect on test strains (Table 4.4), was the most variable trait examined. Inhibition of *Bacillus subtilis* was split almost evenly, and was significantly decreased in isolates from Willsboro, NY (Fisher's one-tailed exact test, p-value = 0.0003). Data is missing because of variable results or multiple test failures (contamination or overgrowth from spore dispersal during overlay application). This test is being repeated for isolates which still have missing data. 8.6% inhibited growth of *E. coli*, 52.8% inhibited *B. subtilis* and 8.6% inhibited *M. smegmatis*.

Biosynthesis and traB *genes*

Biosynthetic gene cluster presence/absence was the most conserved category out of all traits examined (Table 4.5). In these 333 (from 9 gene clusters × 37 isolates) tests, the population only differed from ATCC 33331 in the absence of three of these gene clusters. Average variation from ATCC 33331 in biosynthetic gene clusters is therefore 0.9%, excluding novel clusters in other isolates. It is possible that these primers do not amplify at these three locations due to a polymorphism. Unlike the biosynthetic gene clusters, the presence of plasmid encoded *traB* was highly variable, present in 11 out of 38 isolates. All three isolates missing an NP also contained *traB*.

PKS1 was sequenced using degenerate primers designed for the ketoacyl synthase genes of polyketide synthase type II pathways. The nucleotide diversity of the sequenced region is 0.00276, and there are five segregating sites, none of which lead to changes in the amino acid sequence. There is evidence for recombination in the history of this gene. Hudson's Rmin for this data set is 1, as calculated with DnaSP (27, 28). This means there are all four possible combinations of two polymorphisms at two sites (sites 22 and 178), which indicates the occurrence of either recurrent mutation or, more likely, recombination.

Similarity to other species

Similar 16S rRNA genes were found using a BLAST search of the non-redundant database. Named species with the closest hits over nearly the full gene are found in Table 4.6. Where available, percent identity in the five protein coding genes in other multilocus sequence typing studies and carbon-source utilization data is also in Table 4.6. While the average 16S identity among these isolates is 99.8%, average similarity at MLST genes is only 92.4%.

Table 4.2. Carbon-source utilization for all 38 members of the *pratensis* population.

Isolate #	Glucose	Xylose	Arabinose	Sucrose	Raffinose
1	+	+	+	−	−
2	+	+	+	−	−
3	+	+	+	−	−
4	+	+	+	−	−
5	+	+	+	−	−
6	+	+	+	−	−
7	+	+	+	−	−
8	+	+	+	−	−
9	+	+	+	−	−
10	+	+	+	−	−
11	+	+	+	+	−
12	+	+	+	−	−
13	+	+	+	−	−
14	+	+	+	−	−
15	+	+	+	−	−
16	+	+	+	−	−
17	+	+	+	−	−
18	+	+	+	−	−
19	+	+	+	−	−
20	+	+	+	−	−
21	+	+	+	−	−
22	+	+	+	−	−
23	+	+	+	−	−
24	+	+	+	−	−
25	+	+	+	−	−
26	+	+	+	−	−
27	+	+	+	−	−
28	+	+	+	−	−
29	+	+	+	−	−
30	+	+	+	−	−
31	+	+	+	−	−
32	+	+	+	−	−
33	+	+	+	−	−
34	+	+	+	−	−
35	+	+	+	+	−
36	+	+	+	−	−
37	+	+	+	−	−
pratensis	+	+	+	−	−

Table 4.2 continued

Isolate #	Rhamnose	Mannitol	Inositol	Fructose	Cellobiose
1	+	+	−	+	+
2	+	+	−	+	+
3	+	+	−	+	+
4	+	+	−	+	+
5	+	+	−	+	+
6	+	+	−	+	+
7	+	+	−	+	+
8	+	−	−	+	+
9	+	+	−	+	+
10	+	+	−	+	+
11	+	+	−	+	+
12	+	+	−	+	+
13	+	+	−	−	+
14	+	+	−	+	+
15	+	+	−	+	+
16	+	+	−	+	+
17	+	+	−	−	+
18	+	+	−	+	+
19	+	+	−	+	+
20	+	+	−	+	+
21	+	+	−	+	+
22	+	+	−	+	+
23	+	+	−	+	+
24	+	+	−	+	+
25	+	+	−	+	+
26	+	+	−	+	+
27	+	+	−	+	+
28	+	+	−	+	+
29	+	+	−	+	+
30	+	+	−	+	+
31	+	+	−	−	+
32	+	+	−	+	+
33	+	+	+	+	+
34	+	+	−	+	+
35	+	+	−	+	+
36	+	+	−	+	+
37	+	+	+	+	+
pratensis	+	+	−	+	+

Table 4.3. Antibiotic resistance for the 38 strains of *S. flavogriseus* phylogroup *pratensis*. Resistance (R) and Sensitivity (S) are indicated for penicillin (P), kanamycin (K), chloramphenicol (C), tetracycline (TE), ampicillin (AM), streptomycin (S), erythromycin (E) and oxytetracycline (T)

Isolate #	P	K	C	TE	AM	S	E	T
1	R	S	R	S	R	S	R	S
2	R	S	S	R	R	S	S	R
3	R	S	R	R	R	S	S	R
4	R	S	S	S	R	S	S	R
5	R	S	R	R	R	S	S	R
6	R	S	S	R	R	S	S	R
7	R	S	S	S	R	S	S	R
8	R	S	S	R	R	S	S	S
9	R	S	S	R	R	S	S	R
10	R	S	R	S	R	S	S	R
11	R	S	R	R	R	S	S	R
12	R	S	S	R	R	S	S	S
13	R	S	S	R	R	S	S	R
14	R	S	S	R	R	S	S	R
15	R	S	S	R	R	S	S	R
16	R	S	R	R	R	S	S	R
17	R	S	R	R	R	S	S	S
18	R	S	R	R	R	S	S	R
19	R	S	S	R	R	S	S	R
20	R	S	S	R	R	S	S	R
21	R	S	S	R	R	S	S	R
22	R	S	S	R	R	S	S	R
23	R	S	S	R	R	S	S	R
24	R	S	S	R	R	S	S	R
25	R	S	S	S	R	S	S	R
26	R	S	S	R	R	S	S	R
27	R	S	S	R	R	S	S	R
28	R	S	S	R	R	S	S	R
29	R	S	S	R	R	S	S	R
30	R	S	S	S	R	S	S	R
31	R	S	S	R	R	S	S	R
32	R	S	S	S	R	S	S	S
33	R	S	S	S	R	S	S	R
34	R	S	S	S	R	S	S	R
35	R	S	S	R	R	S	S	R
36	R	S	S	R	R	S	S	R
37	R	S	S	S	R	S	S	R
pratensis	R	S	R	S	R	S	S	R

Table 4.4. Antibiosis of *S. flavogriseus* phylogroup *pratensis* on *E. coli*, *B. subtilis*, and *M. smegmatis*. (+) indicates inhibition of growth, (−) indicates no effect on growth. Location is given because *B. subtilis* antibiosis is significantly different by site.

Isolate #	*E. coli*	*B. subtilis*	*M. smegmatis*	Location
1	−	−	+	Caldwell Field, Ithaca, NY
2	−	+	−	Chazy, NY
3	−	+	−	Caldwell Field, Ithaca, NY
4	−	−	−	Grand Rapids, MI
5	−	−	−	Grand Rapids, MI
6	−	+	−	Grand Rapids, MI
7	−	+	−	Grand Rapids, MI
8	−	+	−	Grand Rapids, MI
9	−	+	+	Harford, NY
10	−	+	−	Harford, NY
11	−	+	−	Harford, NY
12	−	+	−	Mitchell Street, Ithaca, NY
13	−	+	−	Mitchell Street, Ithaca, NY
14	−	+	−	Mitchell Street, Ithaca, NY
15	−	+	−	Mitchell Street, Ithaca, NY
16	+	+	−	Mitchell Street, Ithaca, NY
17	−	+	−	Mitchell Street, Ithaca, NY
18	−	+	−	Mitchell Street, Ithaca, NY
19	−	+	−	Willsboro, NY
20	−	+	−	Willsboro, NY
21	−	−	−	Willsboro, NY
22	−	−	−	Willsboro, NY
23	−	−	−	Willsboro, NY
24	+	−	+	Willsboro, NY
25	−	−	−	Willsboro, NY
26	−	−	−	Willsboro, NY
27	−	−	−	Willsboro, NY
28	−	−	−	Willsboro, NY
29	+	+	−	Willsboro, NY
30	−	−	−	Willsboro, NY
31	M	+	M	Willsboro, NY
32	−	−	−	Willsboro, NY
33	−	−	−	Willsboro, NY
34	−	−	−	Willsboro, NY
35	−	−	−	Willsboro, NY
36	M	M	M	Willsboro, NY
37	M	M	M	Willsboro, NY
pratensis	−	−	−	Laval, Quebec

Table 4.5. Presence or absence of the plasmid encoded *traB* and nine biosynthetic gene clusters.

Isolate #	*traB*	np1	np2	np3	np4	np5	np6	np7	pks1	pks2
1	−	+	+	+	+	+	+	+	+	+
2	−	+	+	+	+	+	+	+	+	+
3	−	+	+	+	+	+	+	+	+	+
4	−	+	+	+	+	+	+	+	+	+
5	+	+	+	+	+	+	+	+	+	+
6	−	+	+	+	+	+	+	+	+	+
7	+	+	+	+	+	+	+	+	+	+
8	+	+	+	+	+	+	+	+	+	+
9	−	+	+	+	+	+	+	+	+	+
10	−	+	+	+	+	+	+	+	+	+
11	−	+	+	+	+	+	+	+	+	+
12	−	+	+	+	+	+	+	+	+	+
13	+	+	+	+	+	+	+	+	+	+
14	+	+	−	+	+	+	+	+	+	+
15	−	+	+	+	+	+	+	+	+	+
16	−	+	+	+	+	+	+	+	+	+
17	−	+	+	+	+	+	+	+	+	+
18	−	+	+	+	+	+	+	+	+	+
19	−	+	+	+	+	+	+	+	+	+
20	+	+	+	+	+	+	+	+	+	+
21	+	+	+	+	−	+	+	+	+	+
22	−	+	+	+	+	+	+	+	+	+
23	−	+	+	+	+	+	+	+	+	+
24	−	+	+	+	+	+	+	+	+	+
25	−	+	+	+	+	+	+	+	+	+
26	−	+	+	+	+	+	+	+	+	+
27	−	+	+	+	+	+	+	+	+	+
28	−	+	+	+	+	+	+	+	+	+
29	−	+	+	+	+	+	+	+	+	+
30	−	+	+	+	+	+	+	+	+	+
31	−	+	+	+	+	+	+	+	+	+
32	+	+	+	+	+	+	+	+	+	+
33	−	+	+	+	+	+	+	+	+	+
34	−	+	+	+	+	+	+	+	+	+
35	+	+	+	+	+	+	+	+	+	+
36	+	+	+	−	+	+	+	+	+	+
37	−	+	+	+	+	+	+	+	+	+
pratensis	+	+	+	+	+	+	+	+	+	+

Table 4.6. Traits of species with similar 16S rRNA genes. Columns are *Streptomyces* species names, % 16S rRNA gene similarity, % similarity at five MLST genes, culture collection numbers, and carbon-source utilization: glucose (1), xylose (2), arabinose (3), sucrose (4), raffinose (5), rhamnose (6), mannitol (7), inositol (8), and fructose (9). M is missing data. Species names in bold match the *pratensis* sugar profile. Data is from: [1] (10), [2] (29), [3] (30), [4] (31).

Species	16S	MLSA	Culture Collection	1	2	3	4	5	6	7	8	9
pratensis	100	100	ATCC 33331	+	+	+	-	-	+	+	-	+
griseus[1]	99.9	92.6	ISP 5235	+	+	-	-	-	-	+	-	+
luridiscabiei[2]	99.8	93.1	LMG 21390	+	+	+	+	+	+	+	+	+
caviscabies[3]	100.0	92.5	ATCC 51298	+	-	-	-	+	-	-	M	-
flavogriseus[1]	99.7	93.5	ISP 5323	+	+	+	-	-	+	+	-	+
ornatus[1]	99.6	92.3	ISP 5307	+	+	-	-	-	-	+	-	+
globisporus[1]	99.9	90.1	ISP 5199	+	+	+	-	-	+	+	-	+
anulatus[1]	100.0	92.9	ISP 5361	+	+	+	-	-	+	+	-	+
fimicarius[1]	100.0	92.4	ISP 5322	+	+	+	-	-	+	+	-	+
baarnensis[1]	99.4	92.4	ISP 5232	+	+	+	-	-	+	+	-	+
cavourensis[4]	99.9	m	NRRL B-8030	+	+	+	-	-	-	+	-	+
badius[1]	99.9	92.3	ISP 5139	+	+	+	-	-	-	+	-	+
microflavus[1]	99.8	93.0	ISP 5331	+	+	-	-	-	+	+	-	+

Table 4.7. Phenotypic and morphological traits of species with closely related 16S rRNA genes and identical carbon-source utilization profiles. Data from International *Streptomyces* Project (10) except for *S. flavogriseus* phylogroup *pratensis* (18).

species	spore chain	surface	clr aerial	clr reverse	melanoid	pigments
pratensis	rectiflexibiles	smooth	gray	yellow	none	yellow
flavogriseus	rectiflexibiles	smooth	gray	yellow	none	trace yellow
globisporus	rectiflexibiles	smooth	yellow	pale yellow	none	yellow
anulatus	rectiflexibiles[1]	smooth	yellow/white	pale yellow	none	trace yellow
fimicarius	rectiflexibiles	smooth	yellow/white	pale yellow	none	red
baarnensis	rectiflexibiles[2]	smooth	inadequate	yellow	none	none

For strains in bold in Table 4.6, further information from the International *Streptomyces* Project (10) can be found in Table 4.7. The only isolate that completely matches *Streptomyces flavogriseus* phylogroup *pratensis* is *Streptomyces flavogriseus*. However, Table 4.6 shows an average identity at MLSA genes of only 93.5%.

Discussion

Streptomyces flavogriseus phylogroup *pratensis* represents a geographically distributed microbial population whose members have housekeeping genes that share high nucleotide similarity (99.82% average similarity) and have rates of gene exchange sufficient to have alleles at five separated genes in linkage equilibrium (17). We examined these strains to determine whether the high rates of gene exchange we observed in the population manifest as higher levels of phenotypic diversity within the population compared to other bacterial populations.

Our hypothesis was that recombination would provide the potential for high levels intraspecies variability, but that the level of variation would depend upon the trait examined as influenced by selective pressures on the population. Traits used to determine taxonomic standing, such as colony morphology and carbon-source utilization, varied minimally. The most variation was seen in fructose utilization with 8% of strains able to use this trait. Other carbon-sources examined were utilized more uniformly by the population, meaning that <8% or >92% were able to use each carbon source. In contrast, higher levels of intraspecies variability have been documented in other bacteria for the traits we have examined. In strains of *Psuedomonas syringae* pv. *Tomato*, for example, 25.5% of strains are unable to use raffinose (32), mannitol and cellobiose utilization is absent from 40% of *Tetragenococcus halophilus* strains (33), 25% of *Burkholderia ambifaria* strains are unable to use inositol and raffinose (34),

and 38% of *Escherichia coli* strains are unable to use raffinose (35). Only in *Xanthomonas axonopodis* pv. *vignicola*, were we able to find levels of intraspecies phenotypic homogeneity that were similar to those we observed in *S. flavogriseus* phylogroup *pratensis* (36). We did not find, however, enough uniformly reported studies for meaningful comparison in terms of habitat, genome size, recombination rate, etc; we can only report that this population of *Streptomyces flavogriseus* phylogroup *pratensis* appears to show little variation in carbon-source utilization compared to other species of bacteria.

Antibiotic resistance is more variable in this population than carbon-source utilization. For example, tetracycline resistance is highly variable with 29% of isolates sensitive to this antibiotic. Other bacterial species have similar levels of intraspecies variation with respect to antibiotic sensitivity. In *E. coli*, for example, the variability depends upon the strains chosen to compare. Miller and Hartl found streptomycin resistance in22% of isolates within the *E. coli* culture collection (35). However, when studying enterotoxigenic *E. coli* isolated from cases of travelers' diarrhea in Kenya, Shaheen *et al*. found 42% of isolates were resistant to tetracycline (37). For *E. coli* isolated from dairy farms, Scaria *et al*. found 47.2% of isolates resistant to chloramphenicol (38). This intraspecies variation in antibiotic resistance is likely due to the frequency with which resistance genes are associated with mobile genetic elements.

Biosynthetic gene clusters were almost perfectly conserved, as only 3 were not found out of 333 tests. The lack of variation in biosynthetic gene cluster occurrence has important implications for the future of drug discovery research. Because novel biosynthetic diversity exists even within the genomes of bacteria studied for their biosynthetic abilities, genome sequencing is a very promising source of new

secondary metabolite leads. For example, 18 biosynthetic gene clusters were found in the *S. coelicolor* genome sequence whereas only four were found previously in decades of laboratory study (4). Our finding that the nine biosynthetic gene clusters of *S. flavogriseus* phylogroup *pratensis* are very well conserved suggests that delineating gene pools (i.e. defining species and populations) will maximize the efficiency of genomic surveys for biosynthetic gene clusters. What remains to be seen is whether other isolates within this population have more capabilities than those found within the type strain ATCC 33331 alone.

Interestingly, while biosynthetic gene clusters were well conserved, antibiosis was the most variable trait. This could be due to differences in regulation and expression of the genes responsible for this trait rather than gene content differences. Antibiosis was significantly different by site, however, suggesting legitimate differences between the strains. This is also the first hint of geographic structure to the population. The genetic variation underlying this trait is currently being examined with genome sequences for multiple population members.

We have found variation in carbon-source utilization, one set of commonly used tests for determining *Streptomyces* taxonomy, within one species, and high levels of phenotypic and 16S rRNA gene similarity in legitimately different species. This means the current system allows both illegitimate novel species proposals and lumping together multiple species that are legitimately different. These results highlight the need for sequence information in bacterial systematics. Prior to the release of the draft genome of *S. flavogriseus* IAF 45-CD (ATCC 33331), our search for a type strain fitting into our population's gene pool was fruitless. A BLAST search using a *Streptomyces* 16S rRNA gene may return over 100 results of at least 99% identity.

This is too many possibilities for any lab to screen every time a possible novel strain is found. Halting novel species proposals due to high 16S rRNA gene similarity is also counter-productive. At least three named species with good standing have a 16S rRNA gene sequence that is 100% identical to that of *S. flavogriseus* IAF 45-CD, but their average nucleotide identity at the five protein coding genes is only 92.6%, far from the 99.8% identity average amongst *pratensis* population members. An ideal scenario for proposing novel or reviewing current species is to have multiple geographically diverse isolates for each species within the genus which can be used for comparison. Progress towards a genus wide MLST survey is already being made, most notable to date are the studies of Guo *et al*. and Rong *et al* (14-16). Full genome sequencing will provide a more complete understanding of *Streptomyces* evolution and systematic units (e.g. are species boundaries conserved across the whole genome?), and will provide essential details for efficient bioprospecting schemes.

REFERENCES

1. Alanis AJ (2005) Resistance to antibiotics: Are we in the post-antibiotic era? *Arch. Med. Res.* 36(6):697-705.
2. Baltz RH (2008) Renaissance in antibacterial discovery from actinomycetes. *Curr. Opin. Pharmacol.* 8(5):557-563.
3. Payne DJ, Gwynn MN, Holmes DJ, & Pompliano DL (2007) Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Disc.* 6(1):29-40.
4. Bentley SD*, et al.* (2002) Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* 417(6885):141-147.
5. Ohnishi Y*, et al.* (2008) Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. *J. Bacteriol.* 190(11):4050-4060.
6. Ikeda H*, et al.* (2003) Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat. Biotechnol.* 21(5):526-531.
7. Tettelin H*, et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.* 102(39):13950-13955.
8. Lukjancenko O, Wassenaar TM, & Ussery DW (2010) Comparison of 61 Sequenced *Escherichia coli* Genomes. *Microb. Ecol.*
9. Trost B, Haakensen M, Pittet V, Ziola B, & Kusalik A (2010) Analysis and comparison of the pan-genomic properties of sixteen well-characterized bacterial genera. *BMC Microbiol.* 10(1):258.

10. Gottlieb D & Shirling E (1967) Cooperative description of type cultures of Streptomyces. I. The international *Streptomyces* project. *Int. J. Syst. Evol. Microbiol.* 17(4):315.
11. Kuster E (1972) Simple Working Key for the Classification and Identification of Named Taxa Included in the International Streptomyces Project. *Int. J. Syst. Bacteriol.* 22(3):139-148.
12. Williams S*, et al.* (1983) Numerical classification of *Streptomyces* and related genera. *Microbiology* 129(6):1743.
13. Kampfer P, Kroppenstedt R, & Dott W (1991) A numerical classification of the genera *Streptomyces* and *Streptoverticillium* using miniaturized physiological tests. *Microbiology* 137(8):1831.
14. Guo Y, Zheng W, Rong X, & Huang Y (2008) A multilocus phylogeny of the *Streptomyces griseus* 16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. *Int. J. Syst. Evol. Microbiol.* 58(Pt 1):149-159.
15. Rong X, Guo Y, & Huang Y (2009) Proposal to reclassify the *Streptomyces albidoflavus* clade on the basis of multilocus sequence analysis and DNA-DNA hybridization, and taxonomic elucidation of *Streptomyces griseus* subsp. *solvifaciens*. *Syst. Appl. Microbiol.* 32(5):314-322.
16. Rong X & Huang Y (2010) Taxonomic evaluation of the *Streptomyces griseus* clade using multilocus sequence analysis and DNA-DNA hybridization, with proposal to combine 29 species and three subspecies as 11 genomic species. *Int. J. Syst. Evol. Microbiol.* 60(3):696.
17. Doroghazi J & Buckley D (2010) Widespread homologous recombination within and between *Streptomyces* species. *ISME J.* 4:1136-1143.

18. Ishaque M & Kluepfel D (1980) Cellulase complex of a mesophilic Streptomyces strain. *Can. J. Microbiol.* 26(2):183.

19. Shirling E & Gottlieb D (1966) Methods for characterization of *Streptomyces* species. *Int. J. Syst. Bacteriol.* 16(3):313.

20. Nkanga E & Hagedorn C (1978) Detection of Antibiotic-Producing *Streptomyces* Inhabiting Forest Soils. *Antimicrob. Agents Chemother.* 14(1):51.

21. Wawrik B, Kerkhof L, Zylstra G, & Kukor J (2005) Identification of unique type II polyketide synthase genes in soil. *Appl. Environ. Microbiol.* 71(5):2232.

22. Li M, Ung P, Zajkowski J, Garneau-Tsodikova S, & Sherman D (2009) Automated genome mining for natural products. *BMC Bioinformatics* 10(1):185.

23. Bentley SD*, et al.* (2004) SCP1, a 356,023 bp linear plasmid adapted to the ecology and developmental biology of its host, *Streptomyces coelicolor* A3(2). *Mol. Microbiol.* 51(6):1615-1628.

24. Altschul S, Gish W, Miller W, Myers E, & Lipman D (1990) Basic local alignment search tool. *J. Mol. Biol.* 215(3):403-410.

25. Chenna R*, et al.* (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31(13):3497-3500.

26. Felsenstein J (1989) PHYLIP-phylogeny inference package (version 3.2). *Cladistics* 5(1):164-166.

27. Hudson R & Kaplan N (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111(1):147.

28. Rozas J, Sánchez-DelBarrio J, Messeguer X, & Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19(18):2496.

29. Park D*, et al.* (2003) *Streptomyces luridiscabiei* sp. nov., *Streptomyces puniciscabiei* sp. nov. and *Streptomyces niveiscabiei* sp. nov., which cause potato common scab disease in Korea. *Int. J. Syst. Evol. Microbiol.* 53(6):2049.

30. Goyer C, Faucher E, & Beaulieu C (1996) *Streptomyces caviscabies* sp. nov., from deep-pitted lesions in potatoes in Québec, Canada. *Int. J. Syst. Evol. Microbiol.* 46(3):635.

31. Skarbek J & Brady L (1978) *Streptomyces cavourensis* sp. nov.(nom. rev.) and *Streptomyces cavourensis* subsp. *washingtonensis* subsp. nov., a Chromomycin-Producing Subspecies. *Int. J. Syst. Evol. Microbiol.* 28(1):45.

32. Shenge K, Stephan D, Mabagala R, Mortensen C, & Wydra K (2008) Molecular characterization of *Pseudomonas syringae* pv. tomato isolates from Tanzania. *Phytoparasitica* 36(4):338-351.

33. Juste A*, et al.* (2008) Genetic and physiological diversity of *Tetragenococcus halophilus* strains isolated from sugar-and salt-rich environments. *Microbiology* 154(9):2600.

34. Dalmastri C*, et al.* (2003) A rhizospheric *Burkholderia cepacia* complex population: genotypic and phenotypic diversity of *Burkholderia cenocepacia* and *Burkholderia ambifaria*. *FEMS microbiology ecology* 46(2):179-187.

35. Miller R & Hartl D (1986) Biotyping confirms a nearly clonal population structure in *Escherichia coli*. *Evolution* 40(1):1-12.

36. Khatri-Chhetri G, Wydra K, & Rudolph K (2003) Metabolic diversity of *Xanthomonas axonopodis* pv. *vignicola*, causal agent of cowpea bacterial blight and pustule. *European Journal of Plant Pathology* 109(8):851-860.

37. Shaheen H*, et al.* (2003) Phenotypic diversity of enterotoxigenic *Escherichia coli* (ETEC) isolated from cases of travelers' diarrhea in Kenya. *Int. J. Infect. Dis.* 7(1):35-41.

38. Scaria J*, et al.* (2010) Comparison of phenotypic and genotypic antimicrobial profiles in *Eschericia coli* and *Salmonella enterica* from the same dairy cattle farms. *Mol. Cell. Probes*.

CHAPTER 5

STREPTOMYCES POPULATION GENETICS AND COMMUNITY VARIABILITY AT MULTIPLE SPATIAL SCALES

Introduction

Streptomyces are the most speciose genus ever studied. There are 376 species on the Approved Lists of Bacterial Names as of November 7, 2010 (1), down from an all time high of nearly 3,000, some of which were only named in patent literature (2). Patents are commonly blamed as a source of the ballooning of *Streptomyces* species designations (e.g. (3, 4)), because patenting a small molecule was an easier task if it came from a novel species. Intensive, rigorous attempts have been made using phenotypic data to provide a framework for the taxonomy and identification of *Streptomyces*. The first large scale attempt was the International *Streptomyces* Project, carried out in hundreds of labs across the world in the late 1960's (5). Later attempts using physiological tests succeeded in grouping *Streptomyces* species into major and minor species clusters, although the relationship between species clusters and species was, and remains, unclear (6, 7).

Recently, changes to *Streptomyces* species designations have been accumulating due to the application of multilocus sequence analysis (MLSA) to multiple groups within the genus. The first of these studies, published by Guo *et al*., examined the *Streptomyces griseus* 16S rRNA gene clade. They found evidence for splitting some subspecies into different species and combining other species into one. There was also one example of misclassification on the genus level: *Streptomyces flavidofuscus* was found to belong to the genus *Nocardiopsis* (8). In another study with striking results,

Rong *et al*. found that eleven different named species and two subspecies should be combined into one species based on MLSA and DNA-DNA hybridization (DDH) data (9). A more in depth study of the *Streptomyces griseus* clade using both MLSA and DDH found that an MLSA distance of 0.007 (0.7% divergence) corresponds to the 70% DDH cutoff for a species, and recommended the combination of 29 species and three subspecies into 11 genomic species (10). Other genus-wide MLSA surveys have made similar conclusions, frequently coupled with DDH analyses. It is common to find examples of multiple named species that should be lumped together based on low sequence diversity for protein-coding genes (11-14). However, several studies that included strains that were not recognized type strains for species found evidence for proposing novel species groups (13, 15, 16).

There is still not consensus on how bacterial diversity, on any level, is distributed, although there are foundational studies that are beginning to elucidate parts of this complicated topic. There have been reports of isolation by geographic distance (e.g. (17-19)), by environmental distance (e.g. (20, 21)) and a combination of these factors (e.g. (22)). The environmental factor that has received the most attention for affecting community assemblage is pH (23-28). The impact of pH is visible when looking at subsets of the bacterial domain as well. For example, changes in both abundance and community composition of Acidobacteria are best explained by soil pH, but are affected by other habitat variables such as precipitation and organic matter content as well (29). In *Salinispora*, some species appear to be globally distributed, while others are difficult to find outside of a small geographic area (30). Members of this genus also produce species specific secondary metabolites, highlighting the potential importance of understanding the distribution of genetic diversity of industrially important bacteria (31). The conclusion from biogeography studies thus far is that for

any given question the answer probably depends upon the study organism.

One force that affects the phylogenetic (and perhaps geographic) distribution of bacterial diversity is homologous recombination. In the most comprehensive uniform examination of prokaryotic homologous recombination to date, the estimated ratio of the effect of recombination to the effect of mutation (r/m) ranged from 0.02-63.6 (32). The largest proportional variation in r/m within a genus was found in *Bacillus*: the lowest r/m, 0.7, is 35% of the highest r/m of 2.0. This is not much variation compared to the over three orders of magnitude spanned by all of the estimated r/m values (32). Should recombination rate be similar within species in each genus, bacterial systematics could rely upon multiple species criteria based on varying population parameters. A clonal clade within a genus would require different considerations than a highly recombinant group, making species designations more complicated. Recombination could also affect the heterogeneity of genetic diversity within a group. A recombinant population could distribute any newly acquired genes horizontally, while a clonal population would inherit almost all genetic material vertically. A selective sweep would purge almost all diversity within a clonal population, while a recombinant group might only experience purging selection at one locus, leaving diversity across the rest of the genome (see (33) for more details).

We examined *Streptomyces* cultivated from fifteen soil samples from across the United States, spanning Florida to Alaska. Isolates were screened primarily by sequencing *rpoB*. We have found no significant relationship between community composition and either environmental or geographic distance. MLSA data was collected for populations spanning the diversity sampled. Populations ranged from clonal, i.e. no evidence of HR in our data set, to highly recombinant. Over 300 of

2,000 colonies isolated from one 50mg soil sample from Ithaca, NY were screened to find interspecies gene exchange. We found no evidence that species boundaries were crossed in these co-habitating streptomycetes. There was large variation in HR rate between species, however, which may cause complications in applying a uniform set of species criteria to the whole genus.

Materials and Methods

Soil collection

Soils were sampled by recruited volunteers and shipped to Cornell University for analysis. Prior to plating, soils were air dried at room temperature. 50 mg of soil was used for each site, although some sites required multiple platings using different 50 mg samples from the site. 50 mg soil was diluted 1:100 in PBS, shaken or vortexed until all visible clumps were dispersed (1-2 minutes), and 25-50 μL were spread onto glycerol-arginine plates containing cycloheximide and Rose Bengal (34, 35). A balance was struck between giving time for sporulation of streptomycetes and picking colonies before *Rhodococcus* and fungi consumed the plate. This resulted in picking *Streptomyces* colonies to streak on separate plates in the range of one to two weeks after initial plating. Table 5.1 shows a list of sites and the number of isolates from each.

Soil analysis

Organic matter content of soil was measured using the loss on ignition technique recommended in the USDA Soil Survey Laboratory Methods Manual (36). Briefly, about two grams of air-dried soil was placed in dry, previously fired and weighed, room temperature crucibles. Crucibles and soil were dried overnight at 110°C then cooled to room temperature in desiccation jars and weighed again. Crucibles and soil

were then heated at 400°C overnight, cooled in desiccation jars and weighed for the final time. Organic matter percentage was calculated as M_L/M_D, where M_L is mass lost on ignition and M_D is mass of dry soil. The pH of soils was measured using a 1:2 dilution (w/v) of soils in 0.01M $CaCl_2$. Soil particles were dispersed using a glass rod to crush and stir soil in the $CaCl_2$ solution for one minute, followed by a four minute rest, repeated five times. All soils were measured in the same way, using the same equipment.

Climate data

Climate data was collected from the National Climatic Data Center. The closest weather station to the sampling sites were used, unless a small change in distance resulted in more complete data sets. Climate data used was mean August temperature, mean January temperature, mean annual temperature and mean total precipitation. Values were averaged over the available data from January 2000 to the present.

DNA purification and data collection

DNA was purified using a method described previously (37). Cells for DNA purification were grown in 96 deep-well plates with gas permeable seals, allowing 1.5 mL cultures for each isolate. For the first soil samples to be screened, REP-PCR was used to dereplicate isolates using the BOX A1R primer (38). The REP-PCR mix was 25 µl for each isolate, consisting of 2.5 µl 10x AmpliTaq Gold Buffer (Applied Biosystems), 6.7 µl $MgCl_2$, 0.4 µl 100 µM A1R primer, 2.5 µl DMSO, 1 µl BSA, 0.25 µl AmpliTaq Gold (5units/µl), 1 µl of 1:10 diluted DNA preparation as template and 9.4 µl H_2O. Thermal cycler conditions were (step 1) 95°C for 15’00, (step 2) 95°C for 1’00, (step 3) 50°C for 1’00, (step 4) 72°C for 8’00, return to (step 2) for 34 additional cycles. As screening proceeded, this was not found to be worth the cost in time and

money, given the number of isolates that were screened by sequencing *rpoB* due to novelty or uncertain grouping of REP-PCR patterns. Soils numbered 1, 2, 4, 5 and 13 and parts of 7 and 10 were screened with a combination of REP-PCR and *rpoB* sequencing. Soils numbered 3, 6, 8, 9, 11 and 12 and parts of 7 and 10 were screened solely with *rpoB*. Only sites 1-13 were used for the broader biogeography study, and sites 14 and 15 were included for population analysis using MLSA. PCR reactions for MLSA were performed as previously described, except as 12.5 µl reactions (one half volume of all reagents). 7 µl of PCR product was cleaned with 1 µl EXOSAP-IT (Amersham Biosciences), following manufacturer's instructions for thermal cycler temperatures and times. BigDye (Applied Biosystems) reactions for Sanger sequencing were performed in 12.5 µl reactions. Each reaction contained 6.05 µl H2O, 2.5 µl 5x Buffer, 0.5 µl DMSO, 0.5 µl BigDye, 0.5 µl of a dGTP/BigDye 5x buffer equal proportions mix, 0.2 µl of 10 µM primer, and 2.25 µl EXOSAP-IT cleaned PCR product. The thermal cycler program for the BigDye reaction was: (step 1) 96°C for 4'00, (step 2) 96°C for 10", (step 3) 60°C for 3'00, return to (step 2) for 24 additional cycles. The usual 50°C annealing step was omitted and greatly improved sequencing success, probably due to the high G+C content of *Streptomyces*. Edge Biosystems 96-well dye terminator removal column purification and sequencing was performed at the Cornell Life Sciences Core Laboratories Center.

Data analysis

REP-PCR gels were analyzed using BioNumerics version 2.0 (Applied Maths). A combination of colony morphology and REP pattern similarity was used to cluster isolates and determine which deserved further screening. For sequence data, CAP3 (39) interfaced with a Perl script was used to uniformly screen trace files, although products were sequenced in only one direction. Alignments were performed using

ClustalW (40) within BioEdit (41). Other manipulations, such as concatenating

Table 5.1. Sampling site numbers, closest landmarks, and coordinates. Sites 14 and 15 were not used in the multiple site *rpoB* study.

Site #	Isolates	Closest City/Town	Abbrev.	latitude	longitude
1	90	Austin, TX	t	30.20	-97.77
2	86	Brookfield, WI	b	43.06	-88.13
3	95	Manley Hot Springs, AK	man	63.87	-149.02
4	92	Palo Alto, CA	st	37.43	-122.17
5	82	Uwharrie, NC	uw	35.71	-79.88
6	40	Denali Highway, AK	den	63.22	-147.68
7	95	Fort Pierce, FL	f	27.54	-80.35
8	85	Kennebunk, ME	m	43.40	-70.54
9	90	Starkville, MS	ms	33.46	-88.80
10	95	Astoria, OR	or	46.18	-123.85
11	67	Sun Prairie, WI	sun	43.17	-89.24
12	53	Bothell, WA	w	47.73	-122.24
13	333	Caldwell Field, Ithaca, NY	only #'s	42.45	-76.46
14	27	Charlotte, NC	ch	38.81	-78.26
15	11	Greensboro, NC	gb	36.09	-79.89

sequence files were performed using Perl scripts. Analysis of the biogeography data set was performed primarily within R using the package vegan (42, 43). Species groups were determined by the 0.01 OTU cutoff from DOTUR output using default settings (44). Relative abundance of each species group at each site was used to make a Bray-Curtis distance matrix. Environmental variables were range transformed and used to calculate a Euclidian distance matrix. Isolation by geographic distance was calculated using Perl scripts and analyzed in R. To plot this date, small variations were applied to point locations (jittered) so that abundance of overlapping points could be discerned. Recombination rate was estimated using the pairwise program in LDhat (45), and reported values are the estimated recombination rate divided by 2. Sequences of *rpoB* not generated in this study were taken from the PubMLST database (46) or

from Genbank genomes, both draft and complete. The NeighborNet phylogenetic network in Figure 5.1 was created using Splitstree version 4 (47). The neighbor-joining (NJ) radial tree in Figure 5.2 was created with Dendroscope v2.7.4 (48) using a Nexus file created by Splitstree. The remaining trees, all NJ, were made with Splitstree. Sequences will be made publicly available through deposition at Genbank. Sequences published by Laskaris *et al*. were retrieved from Genbank and analyzed in the same manner as our own data sets. The clades mentioned include the following isolates: *S. griseus* (CB162, CB163, CR13, DSM 40236, DSM 40627, DSM 40653, DSM 40654, DSM 40657, DSM 40658, DSM 40659, DSM 40660, DSM 40670, DSM 40759, DSM 40855, DSM 40878, and Z34), *S. violaceoruber* (1326 *S. lividans*, BTG 4-723, BTG 4-738, BTG 4-758, BTG 4-759, BTG 6-708, BTG 6-715, BTG 717I, BTG 723I, DSM 40049, DSM 40233, DSM 40419, DSM 40421, DSM 40438, M110 *S. coelicolor*, NRRL B-12000 and DSM 40783), and *S. albidoflavus* (CB218, DSM 40131, 651, E956, E961, CB141, CB143, E948, E953, CB151, CB152, CB171, CB150, CB149, CB146, CB144 and CB148).

Results

Biogeography of Streptomyces spp.

REP-PCR gels were used to identify clusters within half of the sampling sites (see Materials and Methods for details). Dendrograms of REP-PCR patterns are presented in the Appendix. Representatives of each cluster for each site were chosen for *rpoB* sequencing. For the other half of the sampling sites, *rpoB* sequencing was performed for all isolates. Similarity at *rpoB* was used to group isolates into operational taxonomic units (OTUs) defined at 1% cutoff using furthest neighbor clustering implemented in the program DOTUR (44). This resulted in 904 individuals in 89 OTUs across thirteen sites. This data set will be referred to as the 13 site *rpoB* data set

to improve clarity. Relative abundance of each OTU at each of the 13 sites was organized into a matrix (such that all data for any one site would sum to 1) and was used to create the Bray-Curtis community distance matrix. Organic matter content (%OM) and pH of each soil was measured using air-dried soils; climate data was gathered from the NCDC website (Table 5.2). The environmental distance matrix was created based on a matrix of pH, %OM, average January, August and annual temperatures, and average total annual precipitation for each site.

Table 5.2. Soil properties and local climate variables. January, August and Annual categories are average temperatures in the time frame indicated measured in degrees Centigrade. Precipitation is average annual precipitation in centimeters.

Site #	% OM	pH	January	August	Annual	Precipitation
1	8.14	7.29	11.26	30.13	21.04	86.68
2	9.26	6.79	-8.33	20.69	7.33	88.04
3	9.96	5.00	-24.39	13.07	-3.58	58.22
4	6.26	7.11	9.06	19.69	14.46	37.23
5	5.72	5.60	5.04	25.61	15.54	120.20
6	1.89	6.00	-18.84	10.48	-3.92	61.29
7	5.64	7.09	16.39	27.64	22.87	124.72
8	5.74	4.76	-5.63	16.97	7.14	139.84
9	12.84	6.82	6.17	27.11	17.16	153.29
10	8.23	3.90	6.48	16.23	10.81	163.51
11	6.13	6.59	-6.33	20.93	8.37	92.54
12	11.34	5.31	5.38	19.15	11.36	83.10
13	3.49	5.20	-4.85	20.22	8.17	99.79

A Mantel test found significant correlation between the environmental and community distance matrices. This was not the case when both Alaska sites were not included (Table 5.3). The decreased diversity of both Alaska sites and dominance by the same streptomycete results in a low community distance between the sites, 0.46, compared to the average of 0.92 (see Appendix for distance matrices). The next lowest

community distance measure was 0.68 for Caldwell Field, Ithaca, NY and Brookfield, WI. A low community distance measure, the climate similarity of the sites and the magnitude of difference with the other sites dominates the Mantel test. Removing the Denali Highway and Manley Hot Springs data drops the r statistic from 0.275 to 0.1573, and brings the p-value above 0.05 (Table 5.3). However, canonical correspondence analysis does not provide significant results in either case (Table 5.3).

Table 5.3. Mantel test and canonical correspondence analysis (CCA) p-values for furthest-neighbor OTUs created with a cutoff of 1% (OTU 0.01) or 10% (OTU 0.1) divergence.

	All Sites		Omitting Alaska	
Test	OTU 0.01	OTU 0.1	OTU 0.01	OTU 0.1
Mantel	0.008	0.019	0.146	0.36
CCA	0.73	0.55	0.14	0.26

The Denali Highway soil had both the lowest organic matter content and the fewest species present, as only one OTU at the 0.01 cutoff was recovered in all 40 isolates. The soil with the highest organic matter content, Starkville, MS, was found to have 20 OTUs, the most out of any soil, although different numbers of individuals were sampled from each soil. To test significance of the relationship between organic matter and diversity, 40 isolates were randomly chosen from each site 1,000 times, and the Shannon diversity index was calculated for each and averaged for the site. The relationship between organic matter and these average Shannon diversity index values was statistically significant as tested with Spearman's correlation coefficient (rho = 0.478, p-value = 0.0494). Correlation of organic matter content with the number of OTUs present in these resampled 40 isolates was above a significance cutoff of 0.05 (rho = 0.434, p-value = 0.0691).

Phylogenetic Distribution

The average nucleotide diversity in the 13 site *rpoB* data set with only one representative from each OTU is 0.0861; when all sequences are included the average nucleotide diversity drops to 0.0378. Additional *rpoB* sequences were taken from genomes in Genbank and the PubMLST database (46) and included on the NeighborNet phylogenetic network shown in Figure 5.1. Isolates in the 13 site *rpoB* data set are highlighted in red. The part of the network lacking coverage in our data set is, most noticeably, from *Streptomyces hygroscopicus* to *Streptomyces albofaciens*, including *Streptomyces rimosus*. There may also be significant diversity not present in the sequences available from the sources listed.

Species boundaries

REP-PCR was used to screen 304 streptomycetes isolated from the same 50mg of soil from Caldwell Field. Two genes, *rpoB* and *trpB*, were sequenced for 128 of these isolates. These 128 isolates were chosen to represent all of the diversity present based upon REP-PCR gel patterns and colony morphology. All isolates with unique REP-PCR patterns were screened at these two genes, and isolates for whom colony morphology was not the same as others in their REP-PCR group were also screened. No evidence for an interspecies recombination event within these 128 isolates was found (trees may be found in Appendix on pages 177 and 180).

Four genes, *recA*, *rpoB*, *gyrB*, and *trpB* were sequenced for 93 isolates from the same 50 mg, partially overlapping with the 128 isolates described above (Figure 5.2). These 93 isolates were chosen from the six most common REP-PCR groups and comprise what will be referred to as the Caldwell MLSA data set. Additional isolates were

screened for the less abundant of the six groups by colony morphology and *rpoB* sequences. While the 128 isolates mentioned above were chosen based on diversity, the 93 Caldwell MLSA data set isolates were chosen to represent a random sample of their six populations. Within the Caldwell MLSA data set, no evidence for interspecies recombination was found. That is, no alleles in our data set were shared across population boundaries.

The same four genes were sequenced for 101 additional isolates from 11 different sites, comprising what will be referred to as the multiple site MLSA data set (Figure 5.2). There is evidence for a recent interspecies recombination event within this data set. The four gene trees shown in Figure 5.3 show incongruities in relation to isolates f51, f67, f150, gb14, gb15, st140, and st196 (OTU 27). There are also incompatible sites from a recombination event present within OTU 27 considered along with OTU 1 and 5 (Figure 5.4). The most likely scenario is that *rpoB* was transferred from OTU 1 to OTU 27, although it is possible that this recombination event took place in the past between members of the same population and different alleles have fixed in the resultant species. Evidence for recombination and the scenarios described is based on the pattern of incompatible sites shown in Figure 5.4. This incompatibility matrix shows that the informative sites present in the *rpoB* gene of OTU 27 are a product of either recurrent mutation or, more likely, recombination.

Species delineation

Circles are placed on the branches of the radial tree in Figure 5.2 to represent our best estimate of species boundaries. How deep the circle is placed on the branch is arbitrary, but we consider the sequences on the tips past each circle to be part of the same species. Singlets on this tree are considered different species based on the

average intraspecies diversity (0.0025) and the distance between the singlet and its closest relative. The most important of these species delineations splits OTU 5 into two parts. The nucleotide diversity of groups OTU 5.1 and 5.2 is 0.0015 and 0.0016, respectively. The distance between the two most closely related isolates from each group is 0.0116. Further justification is provided in the discussion.

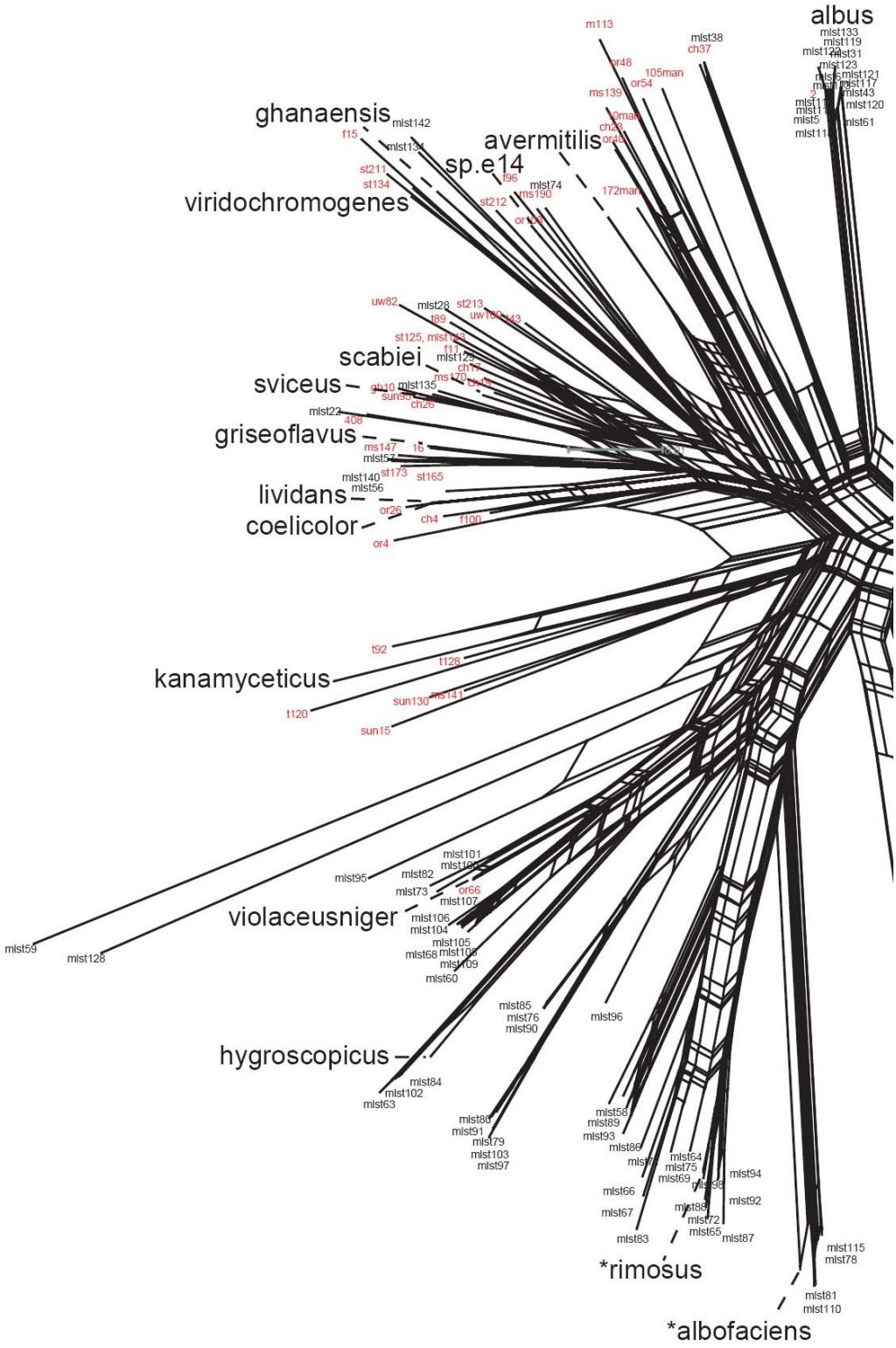
Isolation by distance

The 13 site *rpoB* data set and all of the MLSA concatenated sequences were used to look for isolation by distance. Nucleotide distance between every pair of isolates within each OTU or species was found to weakly correlate with the geographic distance between them, as tested with linear regression. While the r^2 was only 0.038 for the 13 site *rpoB* data set and 0.062 for the combined Caldwell and multiple site MLSA data set, both were significant as tested with linear model ANOVA in R (p-value = <2e-16, and 4.05e-07, respectively). This relationship appears to be driven by an increased number of identical sequences coming from the same soil sample (Figure 5.5).

Recombination rate variation

The four genes sequenced for the Caldwell MLSA and multiple site MLSA data sets were concatenated for analysis of HR rate variation between populations. Sequences for all four genes were trimmed to the same length for all but two groups. Trimming decreased available data for some groups but improved uniformity in the analysis. Due to low polymorphism that inhibited analysis of OTUs 1 and 4, all available data for these groups was used after trimming sequences with only their group members in consideration. This resulted in longer concatenated sequences for OTUs 1 and 4 (Table 5.4). A neighbor-joining radial tree for all concatenated MLSA sequences is

Figure 5.1. Phylogenetic distribution of the 13 site *rpoB* data set depicted with a NeighborNet phylogenetic network. Species names in large font are for *rpoB* sequences taken from genome sequencing projects. Strain names that start with "mlst" are represented in the PubMLST database. Names that start with an asterisk are from the PubMLST database and were chosen to represent their phylogenetic cluster. Names in red are for sequences from this study. One representative of each OTU in the 13 site *rpoB* data set was chosen for inclusion in this network.


albus
ghanaensis
avermitilis
sp.e14
viridochromogenes
scabiei
sviceus
griseoflavus
lividans
coelicolor
kanamyceticus
violaceusniger
hygroscopicus
*rimosus
*albofaciens


sp.Mg1
sp.C
clavuligerus
pristinaespiralis
roseosporus
sp.ACT-1
griseus
sp.ACTE
flavogriseus
sp.SPB74
sp.SPB78
sp.SA3_actG
sp.SA3_actF
*cremeus
mlst49
mlst51
m110
mlst112
mlst19
mlst35
mlst30
t168
184
b47
mlst18
mlst12
st85
st151
mlst50
sun58
b150
b46
m114
mlst136
mlst8
mlst36
sun77
186
mlst126
mlst139
mlst40
st150
mlst3
mlst11
mlst147
113
mlst26
mlst150
mlst130
mlst125
mlst7
mlst53
mlst24
t150
mlst9
mlst13
mlst39
mlst145
mlst29
mlst2
131
mlst149
mlst42
mlst148
t65
mlst141
mlst146
mlst55
mlst15
mlst1
mlst25
w38
mlst47
mlst127
mlst131
mlst20
mlst41
137
mlst16
mlst137
t3
sun107
mlst138
mlst10
or42
w1
mlst32
132
f29
or61
277
mlst44
o43
w10
mlst132
m22
mlst124
uw60
mlst144
mlst4
mlst52
or55
196
or114
mlst54
mlst3
or101
mlst62
m23
sun79
mlst38
t79
man176

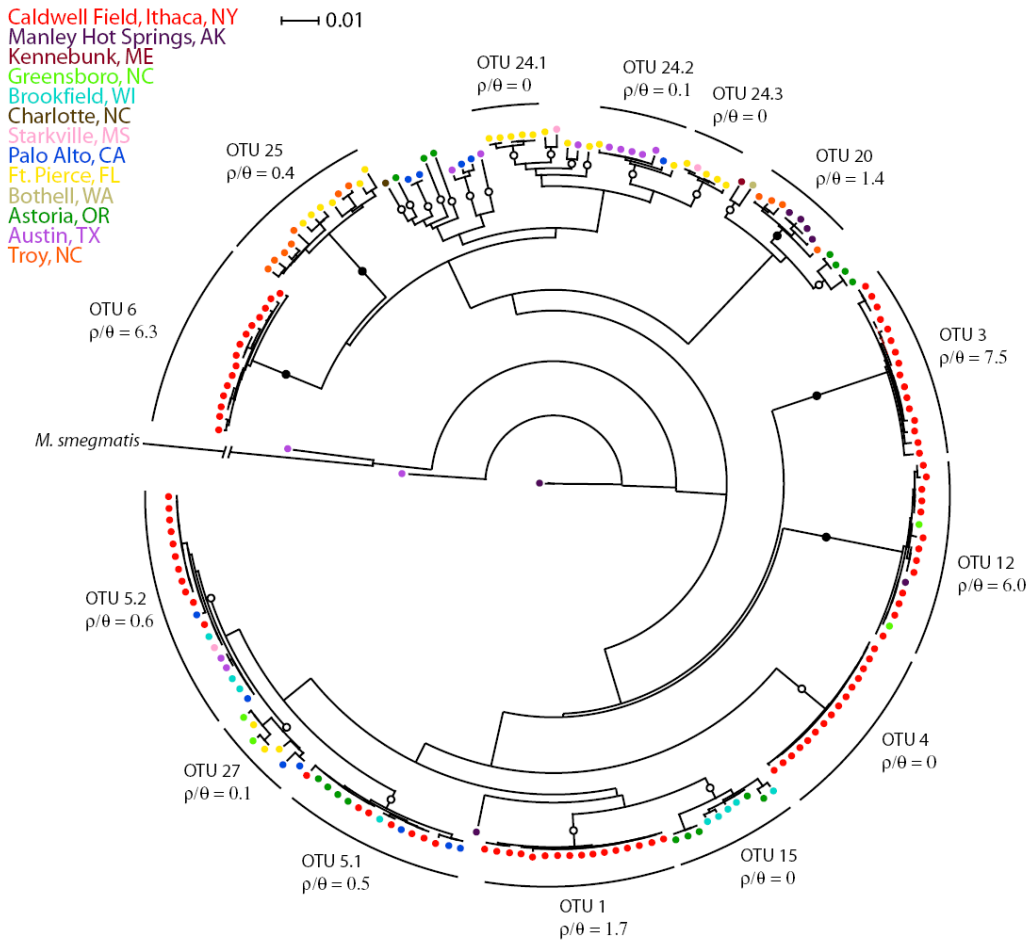
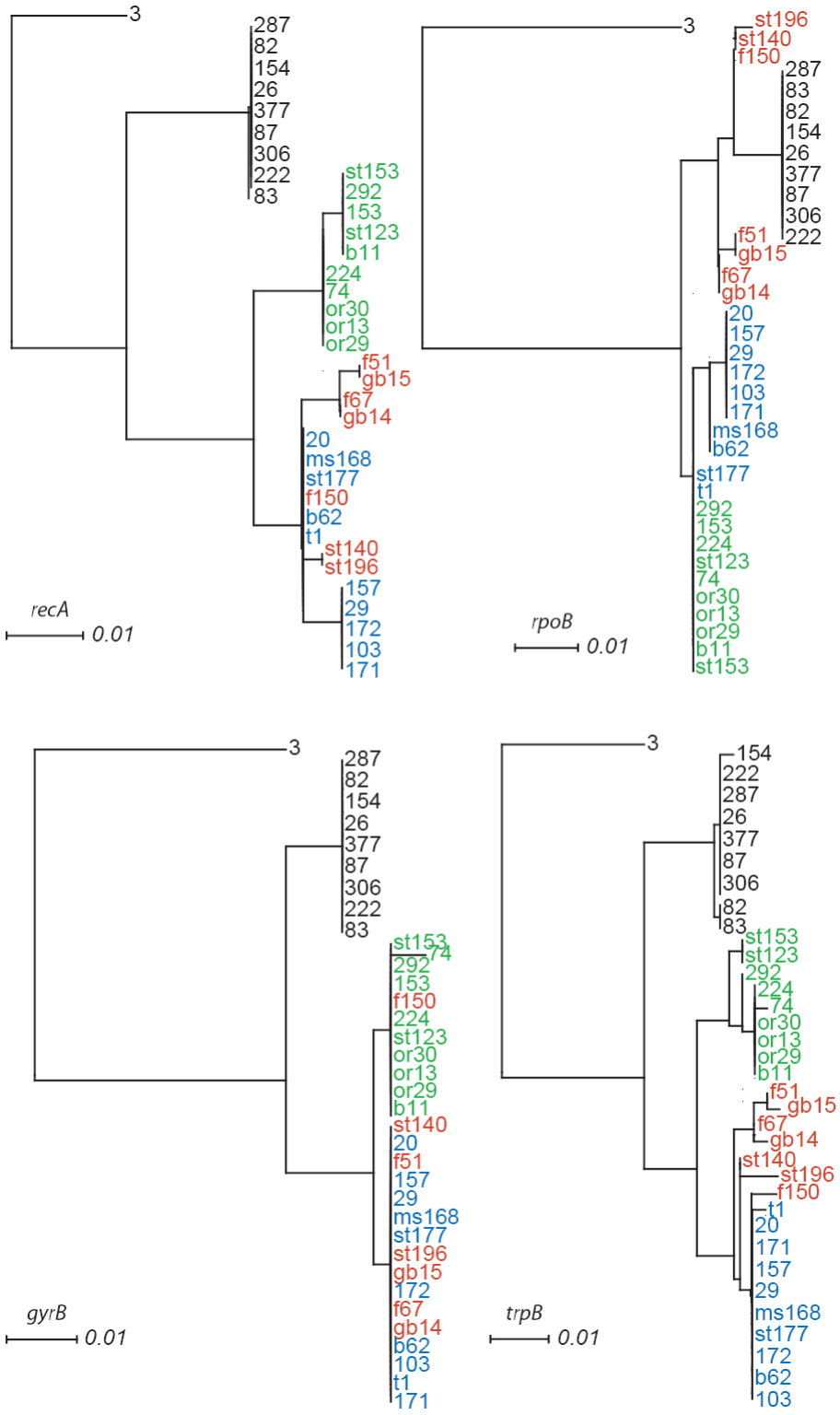


Figure 5.2. Neighbor-joining radial tree of the Caldwell Field and multiple site MLSA data sets. Individuals are represented with a colored circle at branch tips. Circles are colored based on isolation site. Names of sites are colored to correspond with their isolates' circles. The arcs of the divided outer circle delineate species groups used for estimation of population parameters. OTU names and estimated ρ/θ values for each group is provided, as in Table 5.4. Isolates on branch tips past one closed or open circle represent species groups. Closed or open circles are for species groups with or without evidence of recombination, respectively, as determined by a minimum of one recombination event based on incompatible sites (see Table 5.4).

Table 5.4. Population traits for 17 different species from three data sets (Multiple Site MLSA, Caldwell MLSA, and data published by Laskaris *et al*. (49)). Columns are as follows: [a], length of the concatenated sequence analyzed for the species; [b], OTU at 1% cutoff to which the species belongs; [c], number of individuals in the data set; [d], segregating sites in the MLSA sequence for each species; [e], number of unique genotypes within each species; [f], average pairwise nucleotide distance, or nucleotide diversity; [g], Watterson's theta per site (50), or population mutation rate, as estimated with LDhat (45); [h], per-site population recombination rate as estimated with pairwise within LDhat (45); [i], ratio of population recombination to mutation rate; [j], the minimum number of recombination events within the concatenated sequence without recurrent mutation (51), calculated with LDhat (45).

				Multiple Site MLSA					
length[a]	OTU[b]	Indiv.[c]	Seg. Sites[d]	Genotypes[e]	π[f]	θ[g]	ρ[h]	ρ/θ[i]	Rm[j]
1399	25	13	43	13	0.0080	9.9E-03	3.9E-03	0.4	3
1399	24.1	5	3	2	0.0013	1.0E-03	0	0.0	0
1399	24.2	8	9	5	0.0017	2.5E-03	3.6E-04	0.1	0
1399	24.3	5	3	3	0.0010	1.0E-03	0	0.0	0
1399	20	8	11	7	0.0033	3.0E-03	4.3E-03	1.4	3
1399	12	15	6	7	0.0013	1.3E-03	7.9E-03	6.0	1
1399	15	10	19	7	0.0048	4.8E-03	0	0.0	0
1399	5.1	15	7	6	0.0015	1.5E-03	7.1E-04	0.5	0
1399	5.2	19	6	6	0.0016	1.2E-03	7.1E-04	0.6	0
1399	27	7	19	7	0.0055	5.5E-03	7.1E-04	0.1	0
				Caldwell Field MLSA					
1399	5	18	25	8	0.0077	5.2E-03	0	0.0	1
1909	1	16	3	4	0.0004	4.7E-04	7.9E-04	1.7	0
1925	4	15	6	3	0.0004	9.6E-04	0	0.0	0
1399	3	16	12	9	0.0021	2.6E-03	1.9E-02	7.5	2
1399	6	15	7	9	0.0016	1.5E-03	9.6E-03	6.3	1
1399	12	12	6	6	0.0014	1.4E-03	8.6E-03	6.0	1
1399	5.1	7	7	5	0.0018	2.0E-03	0	0.0	0
1399	5.2	11	3	3	0.0006	7.3E-04	0	0.0	0
				Laskaris *et al*. MLSA Data					
3004	*S. griseus*	16	4	3	0.0002	4.0E-04	0	0.0	0
3505	*S. albidoflavus*	17	31	8	0.0028	2.6E-03	7.1E-04	0.3	4
3004	*S. violaceoruber*	17	16	12	0.0017	1.6E-03	0.1	65.5	6

Figure 5.3. Individual neighbor-joining gene trees for OTU 1, 5 and 27 MLSA data. Sequence from an isolate of *S. flavogriseus* phylogroup *pratensis* is used as the outgroup. Isolate names from OTU 5.1 are shown in green, OTU 5.2 are shown in blue, OTU 27 are shown in red, and OTU 1 are shown in black. Some redundant isolates (in terms of novel diversity) from OTUs 1, 5.1 and 5.2 were removed for ease of viewing.


3
287
82
154
26
377
87
306
222
83
st153
292
153
st123
b11
224
74
or30
or13
or29
f51
gb15
f67
gb14
20
ms168
st177
f150
b62
t1
st140
st196
157
29
172
103
171
recA
0.01
3
st196
st140
f150
287
83
82
154
26
377
87
306
222
f51
gb15
f67
gb14
20
157
29
172
103
171
ms168
b62
st177
t1
292
153
224
st123
74
or30
or13
or29
b11
st153
rpoB
0.01
3
287
82
154
26
377
87
306
222
83
st153
74
292
153
f150
224
st123
or30
or13
or29
b11
st140
20
f51
157
29
ms168
st177
st196
gb15
172
f67
gb14
b62
103
t1
171
gyrB
0.01
3
154
222
287
26
377
87
306
82
83
st153
st123
292
224
74
or30
or13
or29
b11
f51
gb15
f67
gb14
st140
st196
f150
t1
20
171
157
29
ms168
st177
172
b62
103
trpB
0.01

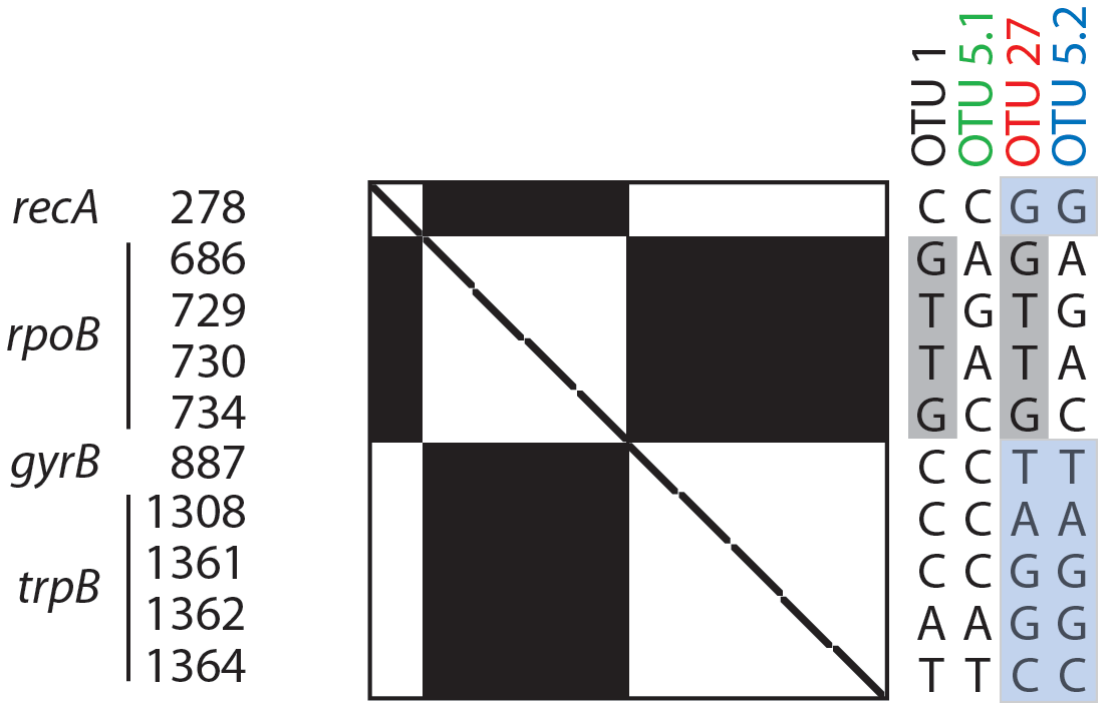
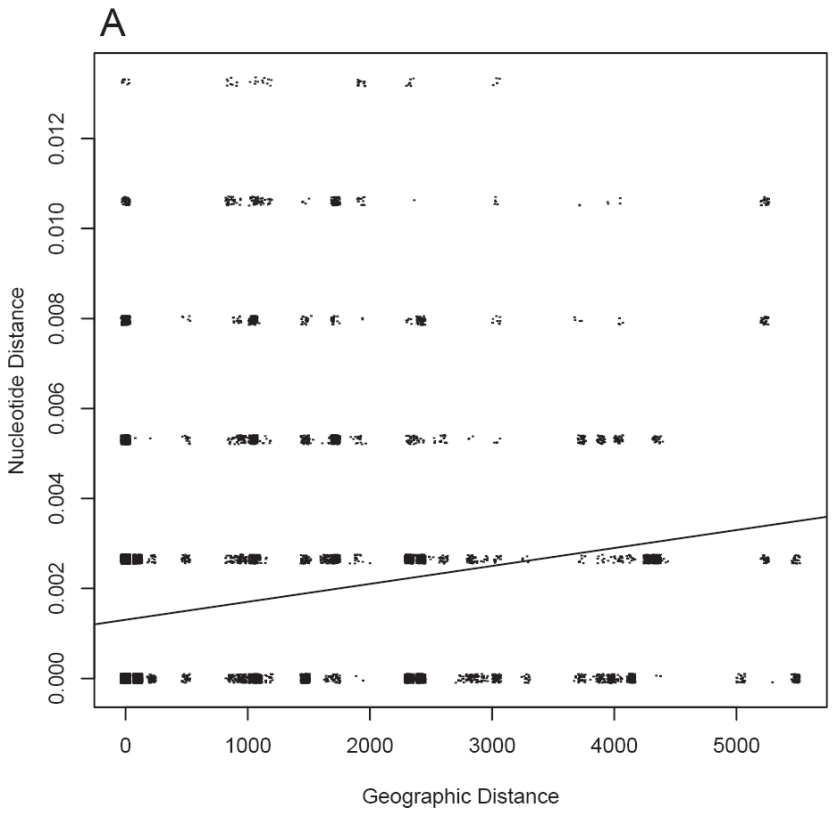
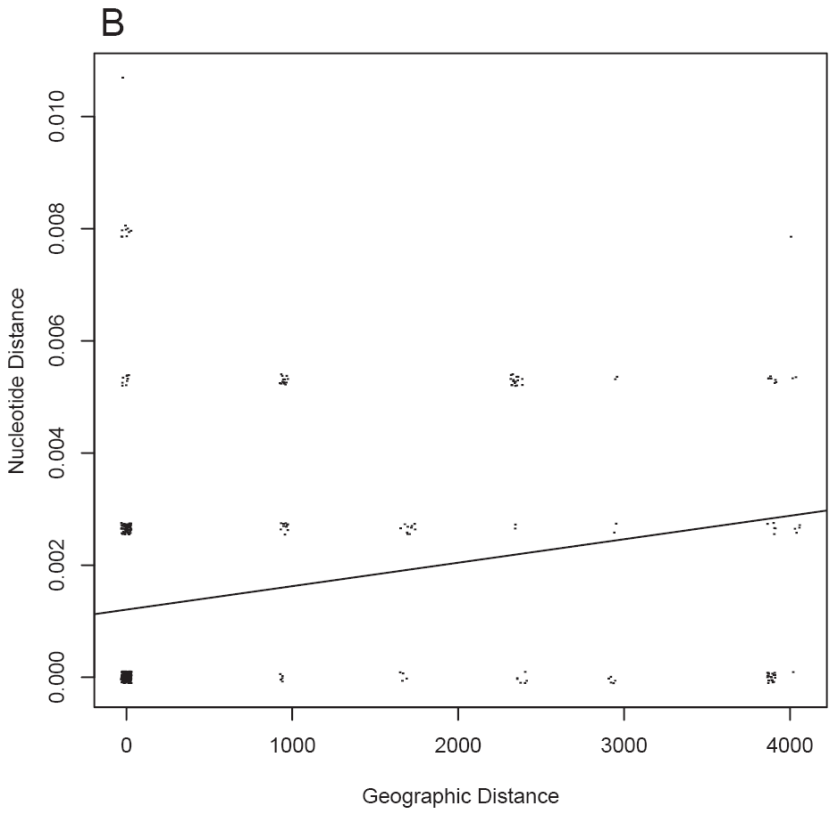


Figure 5.4. Incompatible sites showing recombination involving OTU 27. The incompatible sites matrix was created with the program Reticulate (52). Sites that are not compatible with each other in terms of evolutionary history without recombination are represented by being opposite colors on the symmetric matrix. Sequence sites represented by each row and column are given on the left and are shown next to the gene from which they originate. All informative sites are shown for all four sequences used to make the matrix. Only one representative of each group was used: OTU 1, 222; OTU 5.1, st123; OTU 27, st140; OTU 5.2, st177. The sequence matching OTU 27 is highlighted.

Figure 5.5. Isolation by distance. The 13 site *rpoB* data set (A) and both the Caldwell and multiple sites MLSA data sets (B) were used to calculate isolation by distance. Pairwise comparisons were made between every member of all 89 OTUs and compared with each pairwise geographic distance for the soil sample of origin. Points were given a small random variation (jittered) to make the density of points discernable.


A
Nucleotide Distance
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0
1000
2000
3000
4000
5000
Geographic Distance


B
Nucleotide Distance
0.000
0.002
0.004
0.006
0.008
0.010
0
1000
2000
3000
4000
Geographic Distance

shown in Figure 5.2. Sequence traits for each population are given in Table 5.4. Within the single 50 mg data set, ρ/θ varied from 0 to 7.5. Considering sequences in both MLSA data sets, ρ/θ varied from 0 to 6, although this maximum value is estimated from a data set comprised of individuals predominantly in the Caldwell MLSA data set (Table 5.4). Excluding this value the highest ρ/θ from multiple sites is 1.4 for OTU 20.

Sequences analyzed as part of a study published by Laskaris *et al*. were also used to estimate HR rate in the *S. griseus str* biosynthetic cluster, *S. albidoflavus* clade and the *S. violaceoruber* clade as delineated in the original publication. Some isolates were omitted as they were too divergent to be included as part of the same population (see Materials and Methods for more details). ρ/θ values as estimated for these groups using LDhat are 0, 0.2 and 65.5, respectively (Table 5.4). The MLSA data for these groups consists of partial sequences of seven housekeeping genes, out of which only *rpoB* was also used in the other MLSA data sets.

Discussion

These results have revealed aspects of *Streptomyces* biology that only studying type strains could not. Our primary finding is that there are identifiable species within *Streptomyces*, evidenced by the existence of clusters that were recovered in multiple locations and that do not share the housekeeping genes we have examined with other clusters. This is a different analysis than that performed in Chapter 3, referenced as (37), in which interspecies recombination was found to have a profound impact on *Streptomyces* evolution. Within this study diversity within populations was examined. If interspecies recombination events do, in fact, impact *Streptomyces* evolution, which we still believe is rigorously supported, perhaps such events create novel hybrid

lineages which then diversify and constitute their own species. This would explain how interspecies recombination can have an impact on long term evolution but not be abundant in current populations.

It is difficult to ascertain from four loci what has happened in the history of OTUs 1, 5 and 27. It is possible that the recombination event depicted in Figure 5.4 took place prior to the existence of species boundaries between these groups, where OTUs 1 and 27 fixed one *rpoB* allele and OTU 5 fixed another. OTU 27 is also not strictly more related to either OTU 5.1 or OTU 5.2, as f150 segregates with OTU 5.1 at *gyrB* when the overall trend is a closer relationship to OTU 5.2. The only clear trend with regard to OTU 27 is that it is diverging from OTU 5.

Population genetics theory suggests that a distinct cluster should be considered a species if its closest neighbor is more distant than four times the within cluster nucleotide variation. The nature of random drift makes such a level of diversity within one species unlikely (53, 54). This criterion was used here for delineating species. The highest level of nucleotide diversity within one group we would consider a species is 0.008. This is very close to the value of 0.007 that Rong and Huang have found corresponds to the 70% DNA-DNA hybridization species cutoff within the *S. griseus* clade (10). This explains in part how two legitimate *Streptomyces* species might have identical 16S rRNA genes, as they are generally more conserved than protein coding genes. This is a below average, but still common level of nucleotide diversity for a bacterial population based on the nucleotide diversity of populations in (32). Out of 46 populations for which we have calculated nucleotide diversity using the data sets in (32), 15 have nucleotide diversity below 0.008.

There are two aspects of *Streptomyces* evolution that help to explain the difficulty in producing a systematic taxonomy for the genus. The first is a history of core housekeeping gene exchange that crosses species boundaries (this publication and (37)). Species boundaries do exist, however, suggesting that interspecies recombination events occur at a low frequency and likely produce evolutionarily successful progeny at an even lower frequency. The second trait is the large variation in HR rate across the genus. A clonally reproducing clade would have different traits than a highly recombinant population, and applying a single set of rules to both groups would result in aberrant species designations. Knowledge of these evolutionary details for *Streptomyces* will be essential to the efficient exploration of secondary metabolite production diversity within the genus. The effect of homologous and nonhomologous recombination on population pan-genome diversity is currently unknown. This may affect the number of genomes that must be sequenced for any one species before exhausting the secondary metabolite potential contained therein.

A previous comparison of HR rates across multiple taxa has shown a variation of three orders of magnitude in r/m within bacterial populations studied to date (32). However, estimates for species within the same genus were more similar than would be expected by chance. This does not seem to be the case for *Streptomyces*, however, as our estimates range from clonal ($\rho/\theta = 0$) to recombinant ($\rho/\theta = 7.5$) within the data set we have collected. Analysis of sequence data published by Laskaris *et al*. from the *S. violaceoruber* clade yields much higher estimates of HR rate ($\rho/\theta = 65.5$). For comparison, *Flavobacterium psychrophilum*, the most highly recombinant species in (32) has a ρ/θ value of 28.2 when estimated with LDhat. The program pairwise within LDhat was used for parameter estimation instead of ClonalFrame because we have found it provides more consistent results over a wider range of ρ values. Analysis of

concatenated segments taken from multiple locations on simulated genomes was also found to be relatively accurate. This was found by running both programs on data sets created with known parameter values on a novel bacterial population simulator (data not shown).

The study of the effect of sex/asex on populations is more advanced outside of the microbial realm. Many plants have the capability of reproducing clonally or sexually, and favor one or the other in different situations. While many of the considerations for plants do not apply to bacteria, such as dispersal limitation of asexually produced offspring due to vegetative reproduction as opposed to seeds, some are applicable, such as density dependence. Alien plants that are exploring a new habitat or the plants on the edge of a species' range will have fewer possible mates, and therefore lower rates of sexual reproduction (see (55) for review and meta-analysis). Rare bacteria or a population undergoing a rapid habitat expansion could be influenced in the same way. One repercussion for bacteria that does not exist for plants is that once an expanding population loses a plasmid, or mating factor, it may not regain the plasmid and continue as a clonal population. Silverman concludes his review (55) with the following: "it is now clear that the ecological distribution of more extreme clonality tells us where sex fails, not why it persists." Sexual reproduction, however, has its own set of drawbacks. Alleles that are beneficial in one genomic background may not be in another. Sexual reproduction reshuffles the existing genetic variation, breaking advantageous linkages between co-adapted alleles. In the process, however, new beneficial associations may arise, leading to, for example, the evolution of higher rates of sex in spatially heterogeneous environments (56). The large variation in HR rate we have found between closely related bacterial species offers both a new angle on the investigation of the evolution and ecology of *Streptomyces*, and a new system to

investigate the benefits and cost of gene exchange.

There was no correlation between the community and environmental distance matrices for the majority of the biogeography data set. The one exception to this was the Alaska sites. The communities from the two Alaska soils were by far the most similar, with a community distance of only 0.46, while the next most similar communities were Caldwell Field, Ithaca, NY and Brookfield, WI with a distance of 0.68. It is possible that lower temperatures do select a different subset of the genus, hence the similarity between the Alaskan soils, and that a sampling transect along a gradient of climates would better reveal this result. Our only *a priori* expectation was a pH effect on community composition, as in (23-28). The lack of a correlation between community assemblage and pH could be due to the high pH medium used. This could select a subset of alkiliphilic streptomycetes that would be differentially unaffected, i.e. uniformly affected, by pH variation. The high level of diversity within the genus, the number of OTUs recovered (89 OTUs) and the number of singletons when considering each soil sample individually (52 singletons), suggest that we have undersampled the existing diversity. This is further corroborated by another study that found significantly different communities of streptomycetes within the same 1 m^2 of soil when examining 153 isolates from three locations (57). While undersampling is likely, it could also be possible that the *Streptomyces* community composition of any given site depends upon the previous *Streptomyces* community (or entire biotic) composition of that site more than environmental factors.

Studies that are primarily examining *Streptomyces* biogeography and interaction with their habitat should use overall abundance quantified with culture independent methods (as in (29)) or media at multiple pH values. We have, despite the lack of

correlation between community and environment, discerned that the same *Streptomyces* species can be recovered across the entire range studied. This is likely facilitated by a nearly continuous habitat and the suitability of *Streptomyces* spores for wind dispersal. There was a weak correlation between geographic distance and pairwise nucleotide distance in both *rpoB* and MLSA data sets. This is not the first time that isolation by distance has been found in spore forming bacteria, as it has been previously reported for *Myxococcus xanthus* (18). Populations spanning major migration barriers may, however, show greater isolation by distance, as no significant barriers other than distance were tested within this study. Our hope is that a deeper understanding of the phylogenetic and biogeographic distribution of *Streptomyces* diversity will aid design of systematic and efficient genomic natural products discovery strategies.

REFERENCES

1. Euzéby J (1997) List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. *Int. J. Syst. Evol. Microbiol.* 47(2):590.
2. Trejo W (1970) An evaluation of some concepts and criteria used in the speciation of streptomycetes. *Trans. N. Y. Acad. Sci.* 32:986-997.
3. Kampfer P (2006) The Family Streptomycetaceae, Part I: Taxonomy. 3:538.
4. Anderson A & Wellington E (2001) The taxonomy of *Streptomyces* and related genera. *Int. J. Syst. Evol. Microbiol.* 51(3):797.
5. Gottlieb D & Shirling E (1967) Cooperative description of type cultures of *Streptomyces*. I. The international *streptomyces* project. *Int. J. Syst. Evol. Microbiol.* 17(4):315.
6. Williams S*, et al.* (1983) Numerical classification of *Streptomyces* and related genera. *Microbiology* 129(6):1743.
7. Kampfer P, Kroppenstedt R, & Dott W (1991) A numerical classification of the genera *Streptomyces* and *Streptoverticillium* using miniaturized physiological tests. *Microbiology* 137(8):1831.
8. Guo Y, Zheng W, Rong X, & Huang Y (2008) A multilocus phylogeny of the *Streptomyces griseus* 16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. *Int. J. Syst. Evol. Microbiol.* 58(Pt 1):149-159.
9. Rong X, Guo Y, & Huang Y (2009) Proposal to reclassify the *Streptomyces albidoflavus* clade on the basis of multilocus sequence analysis and DNA-DNA hybridization, and taxonomic elucidation of *Streptomyces griseus* subsp. *solvifaciens*. *Syst. Appl. Microbiol.* 32(5):314-322.

10. Rong X & Huang Y (2010) Taxonomic evaluation of the *Streptomyces griseus* clade using multilocus sequence analysis and DNA-DNA hybridization, with proposal to combine 29 species and three subspecies as 11 genomic species. *Int. J. Syst. Evol. Microbiol.* 60(3):696.
11. Young J, Park D, Shearman H, & Fargier E (2008) A multilocus sequence analysis of the genus *Xanthomonas*. *Syst. Appl. Microbiol.* 31(5):366-377.
12. Naser S*, et al.* (2005) Application of multilocus sequence analysis (MLSA) for rapid identification of *Enterococcus* species based on rpoA and pheS genes. *Microbiology* 151(7):2141.
13. Kotetishvili M*, et al.* (2005) Multilocus sequence typing for studying genetic relationships among *Yersinia* species. *J. Clin. Microbiol.* 43(6):2674.
14. Martens M*, et al.* (2008) Advantages of multilocus sequence analysis for taxonomic studies: a case study using 10 housekeeping genes in the genus *Ensifer* (including former *Sinorhizobium*). *Int. J. Syst. Evol. Microbiol.* 58(1):200.
15. Brady C*, et al.* (2008) Phylogeny and identification of *Pantoea* species associated with plants, humans and the natural environment based on multilocus sequence analysis (MLSA). *Syst. Appl. Microbiol.* 31(6-8):447-460.
16. Norskov-Lauritsen N, Bruun B, & Kilian M (2005) Multilocus sequence phylogenetic study of the genus *Haemophilus* with description of *Haemophilus pittmaniae* sp. nov. *Int. J. Syst. Evol. Microbiol.* 55(1):449.
17. Whitaker RJ, Grogan DW, & Taylor JW (2003) Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science* 301(5635):976-978.
18. Vos M & Velicer GJ (2008) Isolation by distance in the spore-forming soil bacterium *Myxococcus xanthus*. *Curr. Biol.* 18(5):386-391.

19. Cho J & Tiedje J (2000) Biogeography and degree of endemicity of fluorescent *Pseudomonas* strains in soil. *Appl. Environ. Microbiol.* 66(12):5448.

20. Gray N*, et al.* (2007) The biogeographical distribution of closely related freshwater sediment bacteria is determined by environmental selection. *ISME J.* 1(7):596-605.

21. Oakley B, Carbonero F, van der Gast C, Hawkins R, & Purdy K (2010) Evolutionary divergence and biogeography of sympatric niche-differentiated bacterial populations. *ISME J.*

22. Ramette A & Tiedje J (2007) Multiscale responses of microbial life to spatial distance and environmental heterogeneity in a patchy ecosystem. *Proc. Natl. Acad. Sci. U. S. A.*104(8):2761.

23. Fierer N & Jackson RB (2006) The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.* 103(3):626-631.

24. Lauber C, Hamady M, Knight R, & Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* 75(15):5111.

25. Fierer N, Morse J, Berthrong S, Bernhardt E, & Jackson R (2007) Environmental controls on the landscape-scale biogeography of stream bacterial communities. *Ecology* 88(9):2162-2173.

26. Lauber C, Strickland M, Bradford M, & Fierer N (2008) The influence of soil properties on the structure of bacterial and fungal communities across land-use types. *Soil Biol. Biochem.* 40(9):2407-2415.

27. Rousk J*, et al.* (Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J.* 4(10):1340-1351.

28. Nicol G, Leininger S, Schleper C, & Prosser J (2008) The influence of soil pH on the diversity, abundance and transcriptional activity of ammonia oxidizing archaea and bacteria. *Environ. Microbiol.* 10(11):2966-2978.

29. Jones R*, et al.* (2009) A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. *ISME J.* 3(4):442-453.

30. Jensen P & Mafnas C (2006) Biogeography of the marine actinomycete *Salinispora*. *Environ. Microbiol.* 8(11):1881-1888.

31. Jensen P, Williams P, Oh D, Zeigler L, & Fenical W (2007) Species-specific secondary metabolite production in marine actinomycetes of the genus *Salinispora*. *Appl. Environ. Microbiol.* 73(4):1146.

32. Vos M & Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 3(2):199-208.

33. Barraclough T, Birky Jr C, & Burt A (2003) Diversification in sexual and asexual organisms. *Evolution* 57(9):2166-2172.

34. Ottow J (1972) Rose bengal as a selective aid in the isolation of fungi and actinomycetes from natural sources. *Mycologia* 64(2):304-315.

35. El-Nakeeb M & Lechevalier H (1963) Selective isolation of aerobic actinomycetes. *Appl. Environ. Microbiol.* 11(2):75.

36. USDA N (2004) Soil survey laboratory methods manual. *Soil survey investigations report* 42.

37. Doroghazi J & Buckley D (2010) Widespread homologous recombination within and between *Streptomyces* species. *ISME J.* 4:1136-1143.

38. Versalovic J, Schneider M, De Bruijn F, & Lupski J (1994) Genomic fingerprinting of bacteria using repetitive sequence-based polymerase chain reaction. *Methods Mol. Cell. Biol.* 5(1):25-40.

39. Huang X & Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Res.* 9(9):868.

40. Chenna R*, et al.* (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31(13):3497-3500.

41. Hall T (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. pp 95-98.

42. Team RDC (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

43. Jario Oksanen RK, Pierre Legendre, Bob O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Helene Wagner (2009) vegan: Community Ecology Package.

44. Schloss P & Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* 71(3):1501.

45. McVean G, Awadalla P, & Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160(3):1231-1241.

46. Jolley K PubMLST website-Publicly-accessible MLST databases and software.

47. Huson DH & Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23(2):254-267.

48. Huson D*, et al.* (2007) Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 8(1):460.

49. Laskaris P, Tolba S, Calvo-Bado L, & Wellington L (2010) Coevolution of antibiotic production and counter-resistance in soil bacteria. *Environ. Microbiol.* 12(3):783-796.

50. Watterson G (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7(2):256-276.

51. Hudson R & Kaplan N (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111(1):147.

52. Jakobsen IB & Easteal S (1996) A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput. Appl. Biosci.* 12(4):291-295.

53. Birky Jr C, Wolf C, Maughan H, Herbertson L, & Henry E (2005) Speciation and selection without sex. *Rotifera X*:29-45.

54. Birky Jr C, Adams J, Gemmel M, Perry J, & Joly S (2010) Using Population genetic theory and DNA sequences for species detection and identification in asexual organisms. *PLoS ONE* 5(5):e10609.

55. Silvertown J (2008) The evolutionary maintenance of sexual reproduction: evidence from the ecological distribution of asexual reproduction in clonal plants. *Int. J. Plant Sci.* 169(1):157-168.

56. Becks L & Agrawal A (2010) Higher rates of sex evolve in spatially heterogeneous environments. *Nature* 468(7320):89-92.

57. Davelos AL, Xiao K, Samac DA, Martin AP, & Kinkel LL (2004) Spatial variation in Streptomyces genetic composition and diversity in a prairie soil. *Microb. Ecol.* 48(4):601-612.

APPENDIX

Figure A.1. Dendrogram of REP-PCR patterns for streptomycetes isolated from Austin, TX. All isolates shown, except those labeled as “ladder”, are from Austin, TX, whether or not proceeded by “t”.

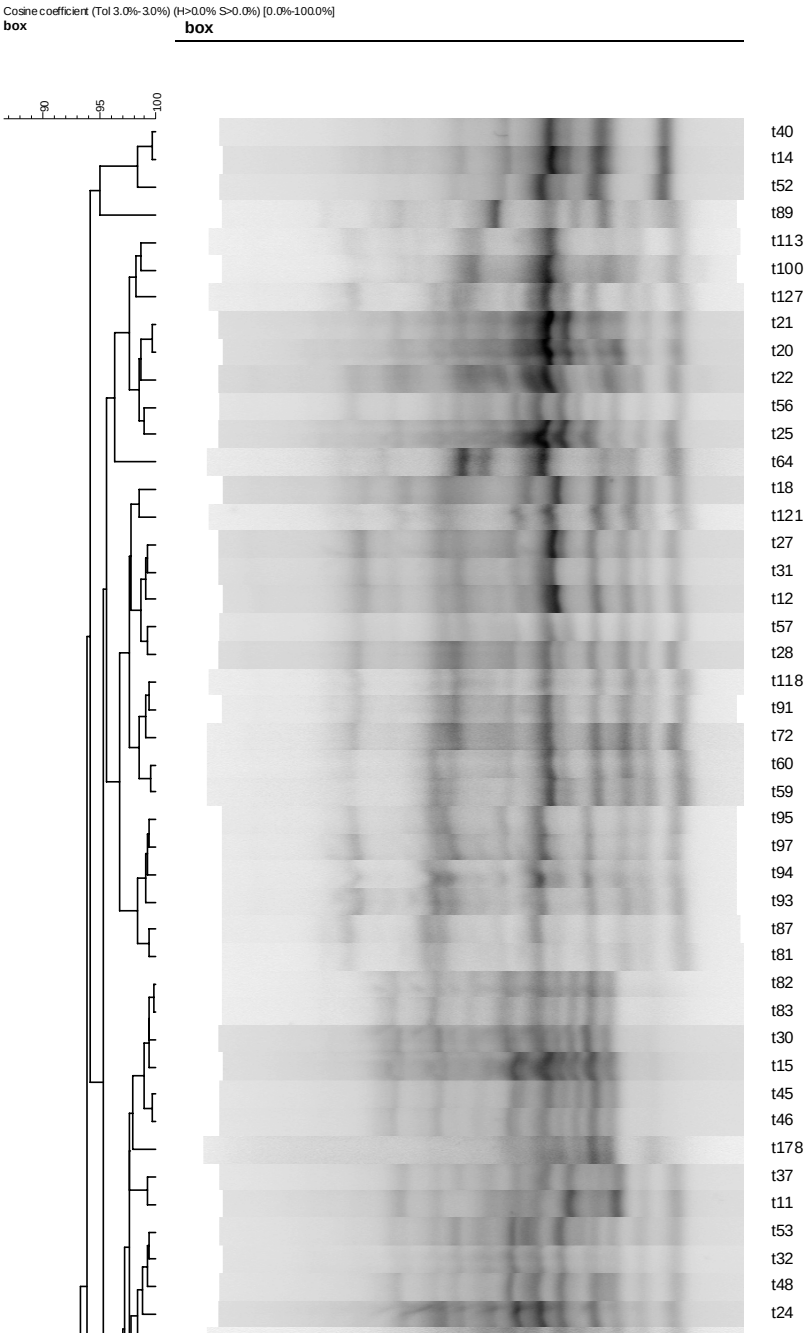

Cosine coefficient (Tol 3.0%-3.0%) (H>0.0% S>0.0%) [0.0%-100.0%]
box
box
90
95
100
t40
t14
t52
t89
t113
t100
t127
t21
t20
t22
t56
t25
t64
t18
t121
t27
t31
t12
t57
t28
t118
t91
t72
t60
t59
t95
t97
t94
t93
t87
t81
t82
t83
t30
t15
t45
t46
t178
t37
t11
t53
t32
t48
t24

Figure A.1. (continued)

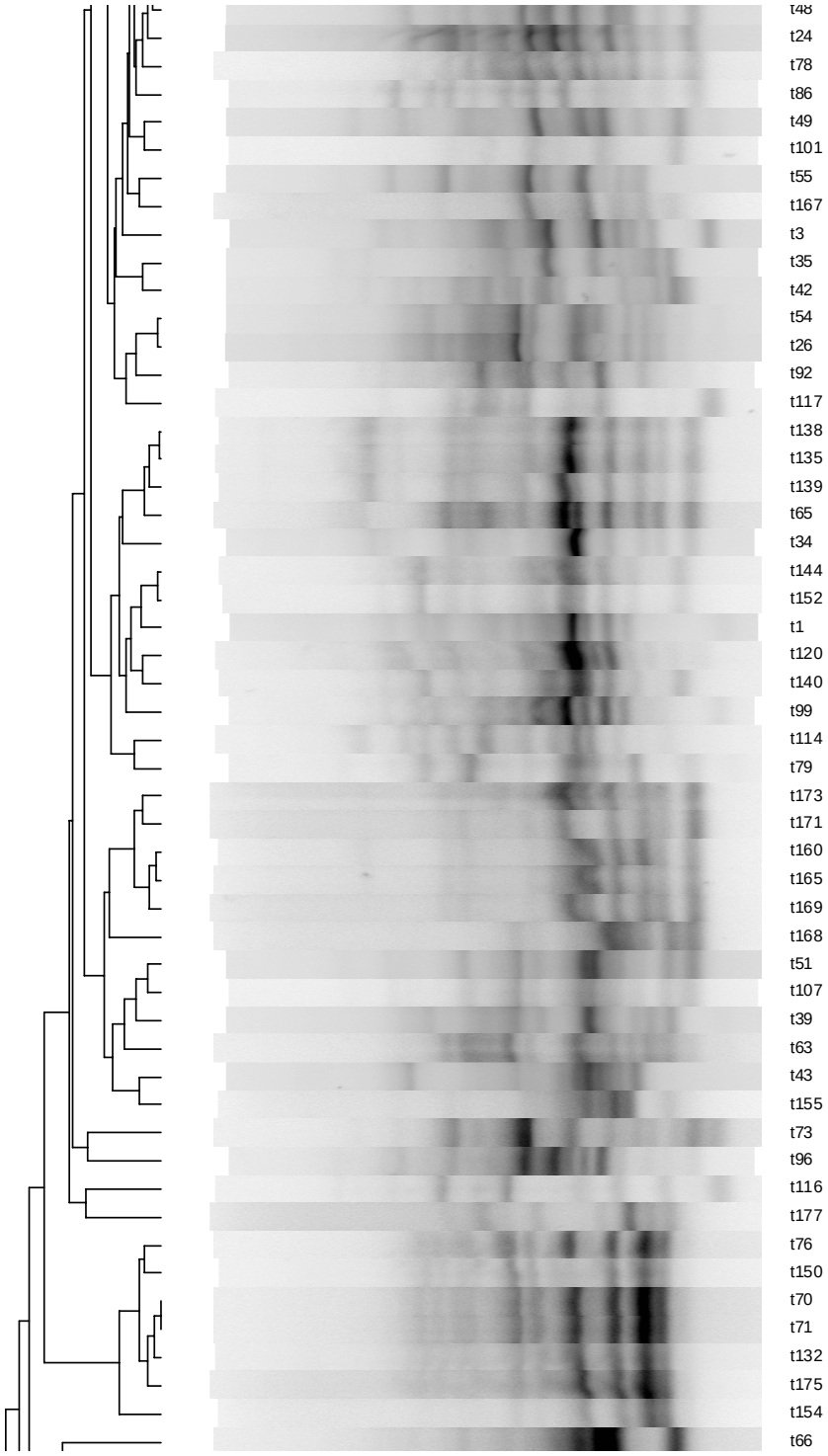


Figure A.1. (continued)

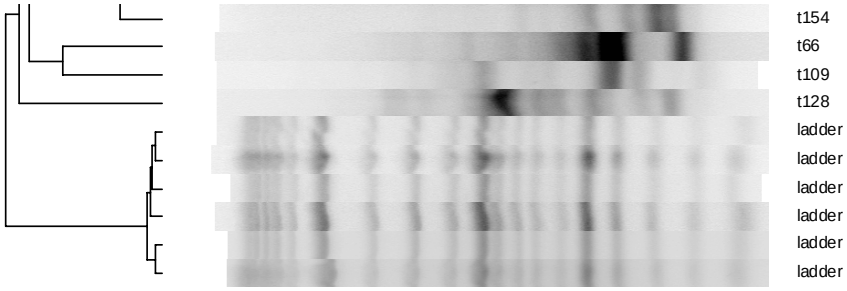


Table A.1. Classification of isolates from Austin, TX based on REP-PCR patterns and *rpoB* sequence data. [a] The isolate number for this soil sample. Throughout Chapter 5 these isolates will be proceeded with a "t". [b] The OTU to which they belong, as determined by the program DOTUR. [c] REP indicates screening via REP-PCR, and those chosen for sequencing at *rpoB* for OTU determination are indicated with "seq".

Isolate[a]	OTU[b]	Screen[c]	Isolate[a]	OTU[b]	Screen[c]
1	OTU 5	REP	76	OTU 10	REP
3	OTU 47	REP, seq	78	OTU 26	REP, seq
11	OTU 26	REP, seq	79	OTU 50	REP, seq
12	OTU 35	REP, seq	81	OTU 35	REP
14	OTU 24	REP, seq	82	OTU 44	REP
15	OTU 44	REP, seq	83	OTU 44	REP
18	OTU 35	REP, seq	86	OTU 35	REP
20	OTU 26	REP, seq	87	OTU 35	REP
21	OTU 26	REP	89	OTU 51	REP, seq
22	OTU 38	REP, seq	91	OTU 35	REP
24	OTU 35	REP	92	OTU 52	REP, seq
25	OTU 35	REP	93	OTU 35	REP
27	OTU 35	REP	94	OTU 35	REP
26	OTU 10	REP, seq	95	OTU 35	REP
28	OTU 35	REP	96	OTU 44	REP, seq
30	OTU 44	REP, seq	97	OTU 35	REP
31	OTU 35	REP	99	OTU 5	REP, seq
32	OTU 26	REP, seq	101	OTU 24	REP, seq
34	OTU 35	REP, seq	107	OTU 46	REP
35	OTU 45	REP, seq	109	OTU 3	REP, seq
37	OTU 26	REP	113	OTU 38	REP
39	OTU 10	REP, seq	118	OTU 35	REP
40	OTU 24	REP	120	OTU 44	REP, seq
43	OTU 48	REP, seq	121	OTU 35	REP
45	OTU 44	REP	127	OTU 35	REP, seq
46	OTU 44	REP	128	OTU 45	REP, seq
48	OTU 26	REP	132	OTU 10	REP
49	OTU 24	REP, seq	135	OTU 35	REP
51	OTU 46	REP, seq	138	OTU 35	REP
52	OTU 24	REP	139	OTU 35	REP
53	OTU 26	REP	140	OTU 50	REP
54	OTU 10	REP, seq	144	OTU 50	REP
55	OTU 49	REP, seq	150	OTU 10	REP
56	OTU 26	REP, seq	152	OTU 50	REP
57	OTU 35	REP	154	OTU 10	REP
59	OTU 35	REP	155	OTU 44	REP, seq
60	OTU 35	REP	160	OTU 35	REP
63	OTU 10	REP, seq	165	OTU 35	REP
64	OTU 38	REP	167	OTU 50	REP
65	OTU 35	REP	168	OTU 46	REP, seq
66	OTU 24	REP, seq	169	OTU 35	REP
70	OTU 10	REP	171	OTU 35	REP
71	OTU 10	REP	173	OTU 35	REP
72	OTU 35	REP	175	OTU 10	REP
73	OTU 38	REP, seq	178	OTU 44	REP, seq

Figure A.2. Dendrogram of REP-PCR patterns for streptomycetes isolated from Brookfield, WI. All isolates shown, except those labeled as “ladder”, are from Brookfield, WI, whether or not proceeded by “b”.

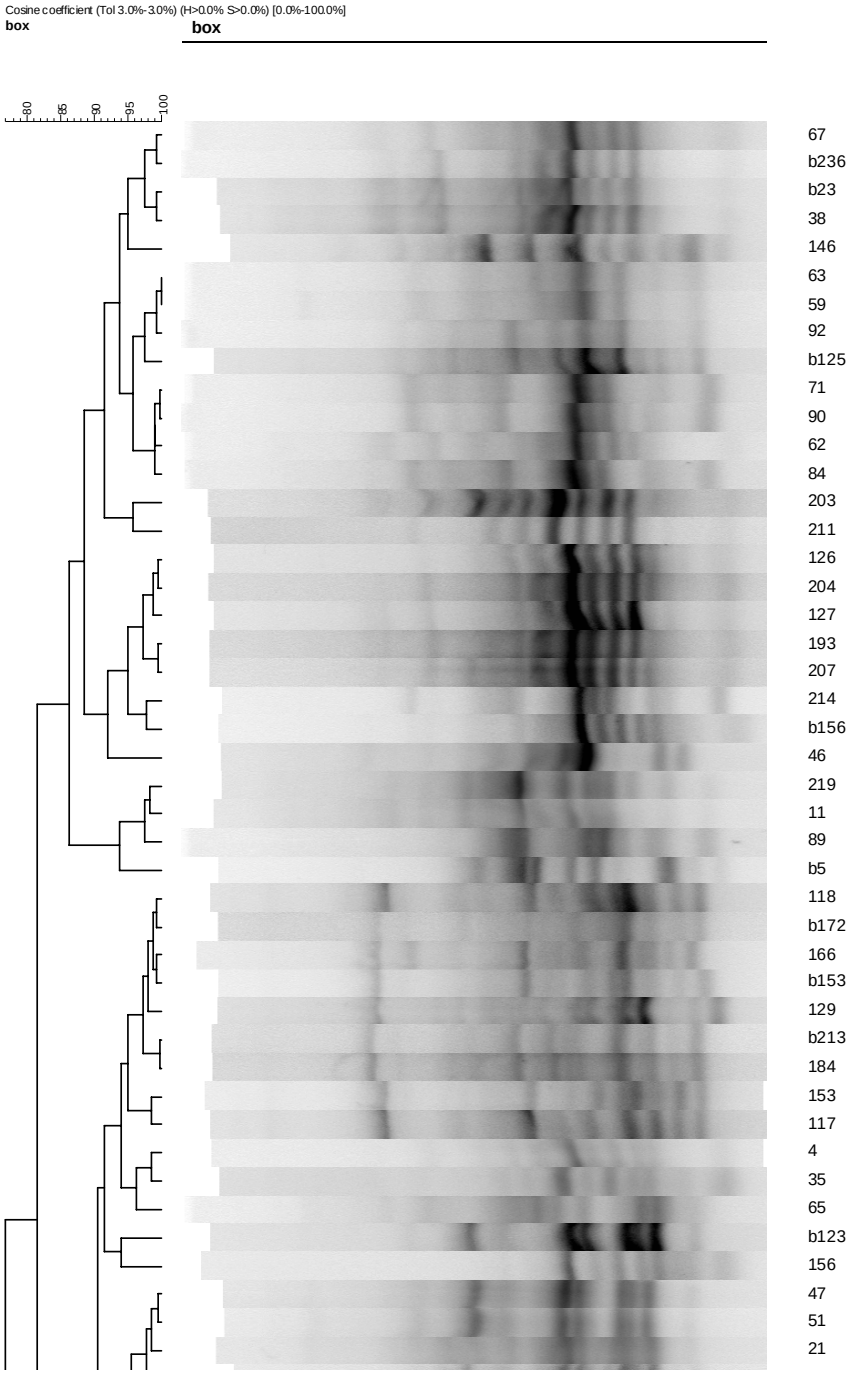

Cosine coefficient (Tol 3.0%-3.0%) (H>0.0% S>0.0%) [0.0%-100.0%]
box
box
80
85
90
95
100
67
b236
b23
38
146
63
59
92
b125
71
90
62
84
203
211
126
204
127
193
207
214
b156
46
219
11
89
b5
118
b172
166
b153
129
b213
184
153
117
4
35
65
b123
156
47
51
21

Figure A.2 (continued)

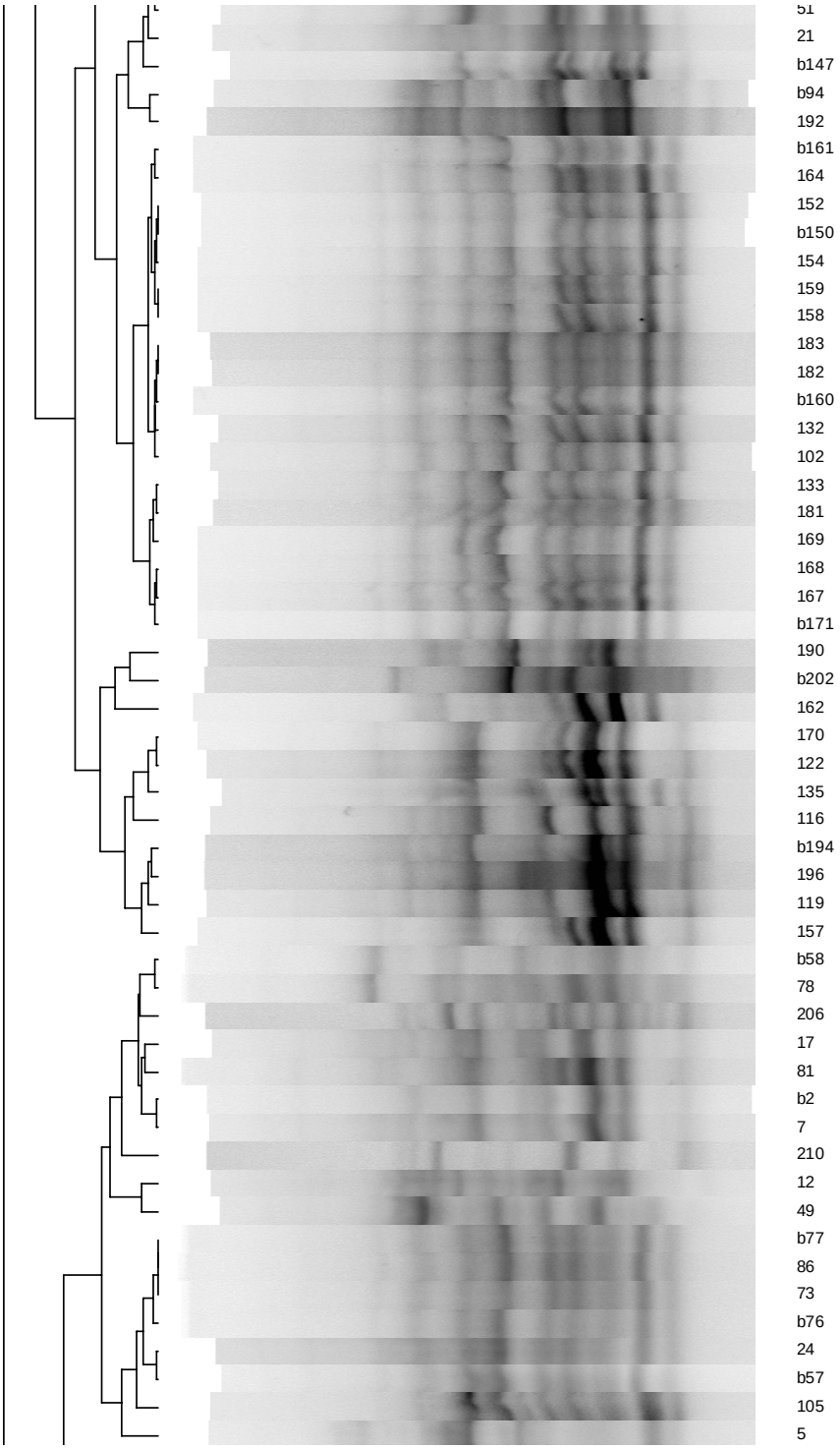


Figure A.2 (continued)

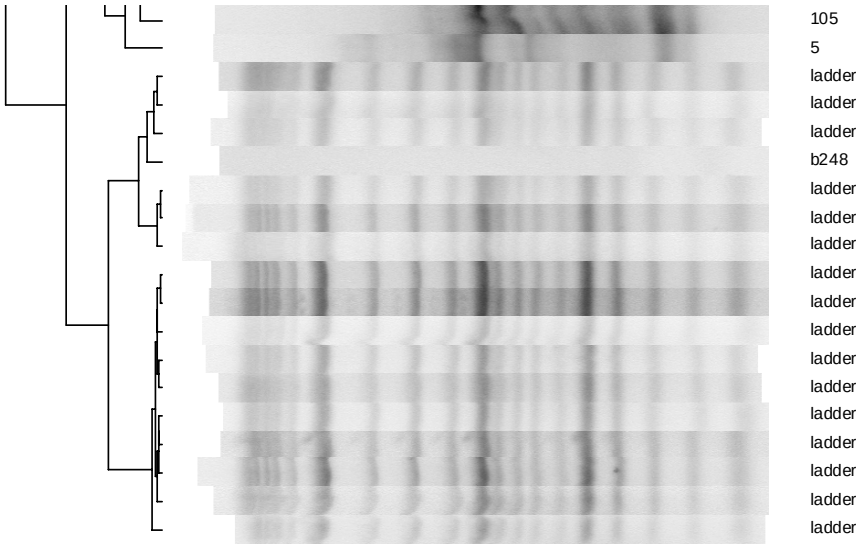


Table A.2. Classification of isolates from Brookfield, WI based on REP-PCR patterns and *rpoB* sequence data. [a] The isolate number for this soil sample. Throughout Chapter 5 these isolates will be proceeded with a "b". [b] The OTU to which they belong, as determined by the program DOTUR. [c] REP indicates screening via REP-PCR, and those chosen for sequencing at *rpoB* for OTU determination are indicated with "seq".

Isolate[a]	OTU[b]	Screen[c]	Isolate[a]	OTU[b]	Screen[c]
2	OTU 3	REP, seq	125	OTU 14	REP
4	OTU 5	REP	126	OTU 5	REP, seq
5	OTU 11	REP, seq	127	OTU 5	REP
7	OTU 3	REP	129	OTU 7	REP
11	OTU 5	REP, seq	132	OTU 14	REP
12	OTU 2	REP, seq	133	OTU 14	REP
17	OTU 3	REP	135	OTU 3	REP
21	OTU 17	REP	147	OTU 17	REP
23	OTU 5	REP	150	OTU 14	REP, seq
24	OTU 14	REP	152	OTU 14	REP
35	OTU 5	REP	153	OTU 7	REP
38	OTU 5	REP	154	OTU 14	REP
46	OTU 16	REP, seq	156	OTU 17	REP, seq
47	OTU 17	REP, seq	157	OTU 3	REP
49	OTU 2	REP, seq	158	OTU 14	REP
51	OTU 17	REP	159	OTU 14	REP
57	OTU 14	REP	160	OTU 14	REP
58	OTU 7	REP, seq	161	OTU 14	REP
59	OTU 14	REP	162	OTU 3	REP
62	OTU 5	REP, seq	164	OTU 14	REP
63	OTU 14	REP, seq	166	OTU 7	REP
65	OTU 17	REP	167	OTU 14	REP
67	OTU 5	REP	168	OTU 14	REP
71	OTU 15	REP, seq	169	OTU 14	REP
73	OTU 14	REP	170	OTU 3	REP
76	OTU 14	REP, seq	171	OTU 14	REP
77	OTU 14	REP	172	OTU 7	REP
78	OTU 7	REP	181	OTU 14	REP
81	OTU 3	REP, seq	182	OTU 14	REP
84	OTU 15	REP	183	OTU 14	REP
86	OTU 14	REP	184	OTU 7	REP
89	OTU 15	REP, seq	192	OTU 2	REP
90	OTU 15	REP	193	OTU 5	REP
92	OTU 14	REP	194	OTU 3	REP
94	OTU 2	REP, seq	196	OTU 3	REP
102	OTU 14	REP	202	OTU 4	REP, seq
105	OTU 14	REP	203	OTU 5	REP, seq
116	OTU 3	REP	204	OTU 5	REP
117	OTU 7	REP, seq	207	OTU 5	REP
118	OTU 7	REP	211	OTU 5	REP
119	OTU 3	REP	213	OTU 7	REP
122	OTU 3	REP	214	OTU 15	REP, seq
123	OTU 17	REP	236	OTU 5	REP

Figure A.3. Dendrogram of REP-PCR patterns for streptomycetes isolated from Palo Alto, CA. All isolates shown, except those labeled as “ladder”, are from Palo Alto, CA, whether or not proceeded by “st”.

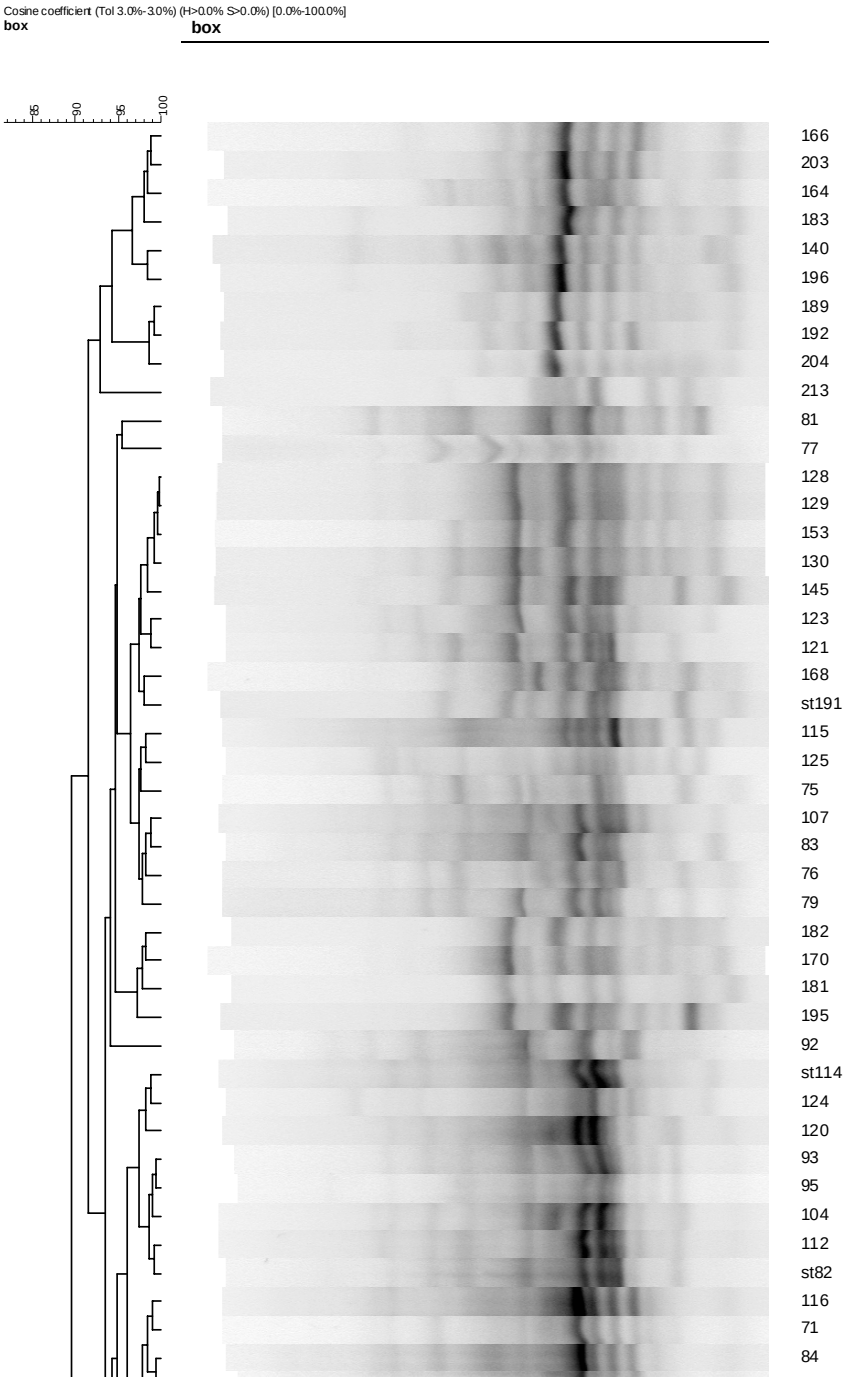

Cosine coefficient (Tol 3.0%-3.0%) (H>0.0% S>0.0%) [0.0%-100.0%]
box
box
85
90
95
100
166
203
164
183
140
196
189
192
204
213
81
77
128
129
153
130
145
123
121
168
st191
115
125
75
107
83
76
79
182
170
181
195
92
st114
124
120
93
95
104
112
st82
116
71
84

Figure A.3. (continued)

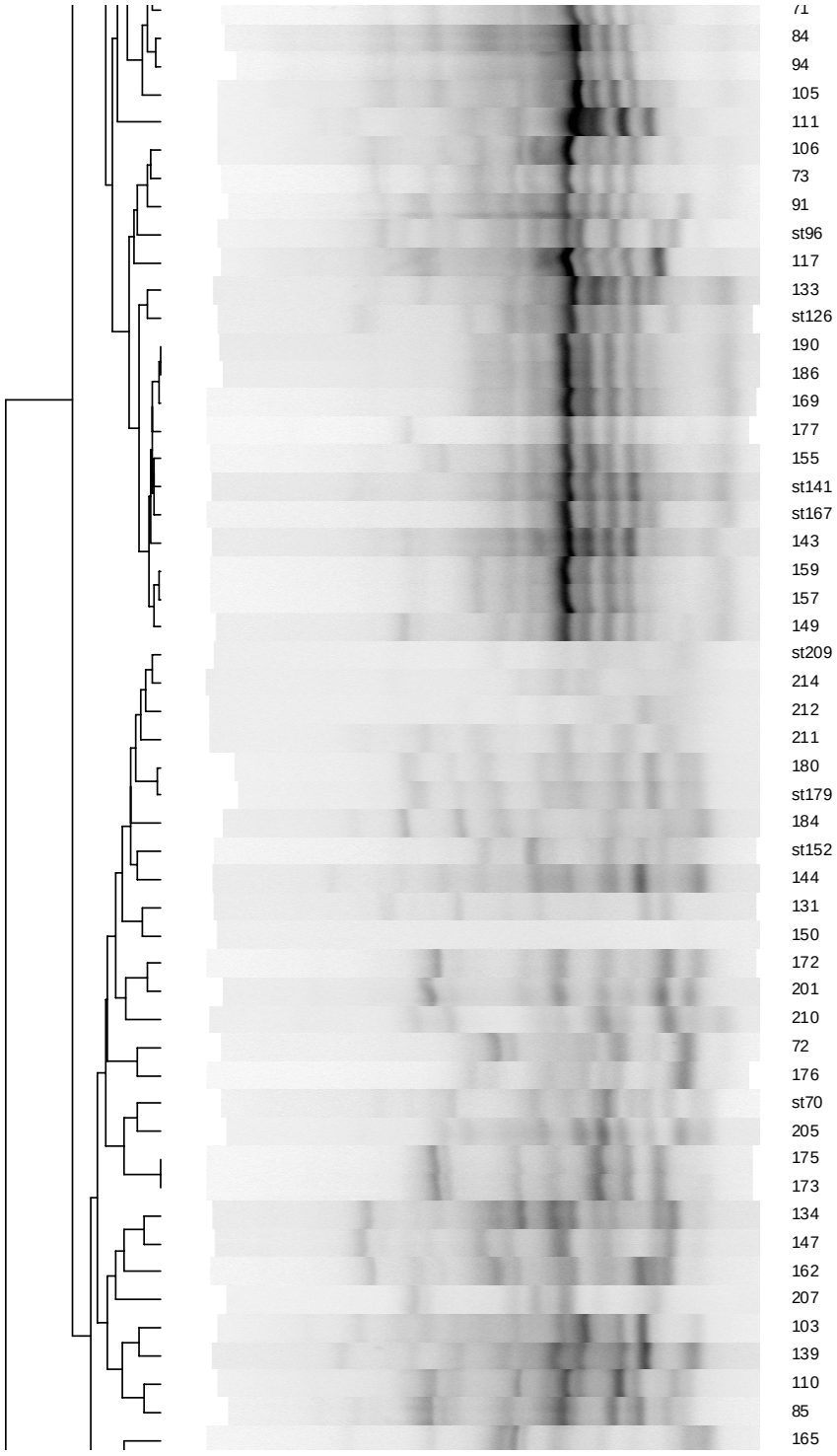


Figure A.3 (continued)

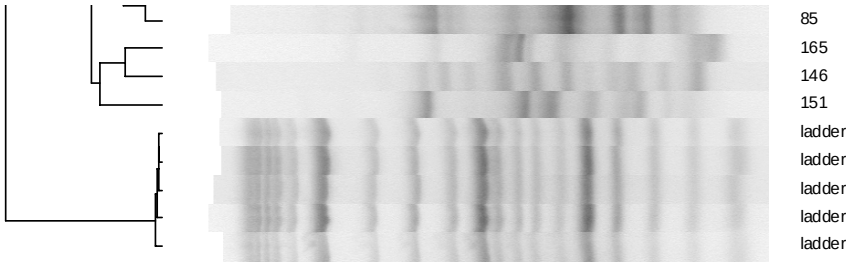


Table A.3. Classification of isolates from Palo Alto, CA based on REP-PCR patterns and *rpoB* sequence data. [a] The isolate number for this soil sample. Throughout Chapter 5 these isolates will be proceeded with the letters "st". [b] The OTU to which they belong, as determined by the program DOTUR. [c] REP indicates screening via REP-PCR, and those chosen for sequencing at *rpoB* for OTU determination are indicated with "seq".

Isolate[a]	OTU[b]	Screen[c]	Isolate[a]	OTU[b]	Screen[c]
70	OTU 34	REP, seq	144	OTU 14	REP
71	OTU 5	REP	145	OTU 5	REP
72	OTU 24	REP, seq	146	OTU 26	REP, seq
73	OTU 5	REP	147	OTU 35	REP
75	OTU 5	REP	149	OTU 5	REP
76	OTU 5	REP	150	OTU 36	REP, seq
77	OTU 36	REP, seq	151	OTU 37	REP, seq
79	OTU 5	REP	153	OTU 5	REP, seq
81	OTU 38	REP, seq	155	OTU 5	REP
82	OTU 36	REP, seq	157	OTU 5	REP
83	OTU 5	REP	159	OTU 5	REP
84	OTU 5	REP	162	OTU 35	REP
85	OTU 40	REP, seq	164	OTU 1	REP
91	OTU 5	REP	165	OTU 38	REP, seq
92	OTU 13	REP, seq	166	OTU 1	REP
93	OTU 36	REP	167	OTU 5	REP
94	OTU 5	REP	168	OTU 5	REP
95	OTU 36	REP	169	OTU 5	REP
96	OTU 35	REP, seq	170	OTU 5	REP, seq
103	OTU 14	REP	172	OTU 23	REP, seq
104	OTU 36	REP	173	OTU 39	REP, seq
105	OTU 5	REP	175	OTU 39	REP
106	OTU 5	REP	176	OTU 24	REP
107	OTU 5	REP	177	OTU 5	REP, seq
110	OTU 40	REP	179	OTU 14	REP
111	OTU 5	REP	180	OTU 14	REP, seq
112	OTU 36	REP	181	OTU 5	REP
114	OTU 36	REP	182	OTU 5	REP
116	OTU 5	REP	183	OTU 1	REP
115	OTU 1	REP, seq	184	OTU 7	REP, seq
117	OTU 5	REP	186	OTU 5	REP
120	OTU 36	REP	189	OTU 1	REP
121	OTU 5	REP	190	OTU 5	REP
123	OTU 5	REP, seq	191	OTU 5	REP
125	OTU 34	REP, seq	192	OTU 1	REP
126	OTU 5	REP	195	OTU 5	REP
128	OTU 5	REP	196	OTU 1	REP, seq
129	OTU 5	REP	201	OTU 23	REP
130	OTU 5	REP	203	OTU 1	REP
131	OTU 34	REP	204	OTU 1	REP
133	OTU 5	REP, seq	205	OTU 1	REP, seq
134	OTU 35	REP, seq	207	OTU 40	REP, seq
139	OTU 14	REP, seq	210	OTU 200	REP, seq
140	OTU 1	REP	211	OTU 41	REP, seq
141	OTU 5	REP	212	OTU 42	REP, seq
143	OTU 5	REP	213	OTU 43	REP, seq

Figure A.4. Dendrogram of REP-PCR patterns for streptomycetes isolated from Ft. Pierce, FL. All isolates shown, except those labeled as “ladder” are from Ft. Pierce, FL, whether or not proceeded by “f”.

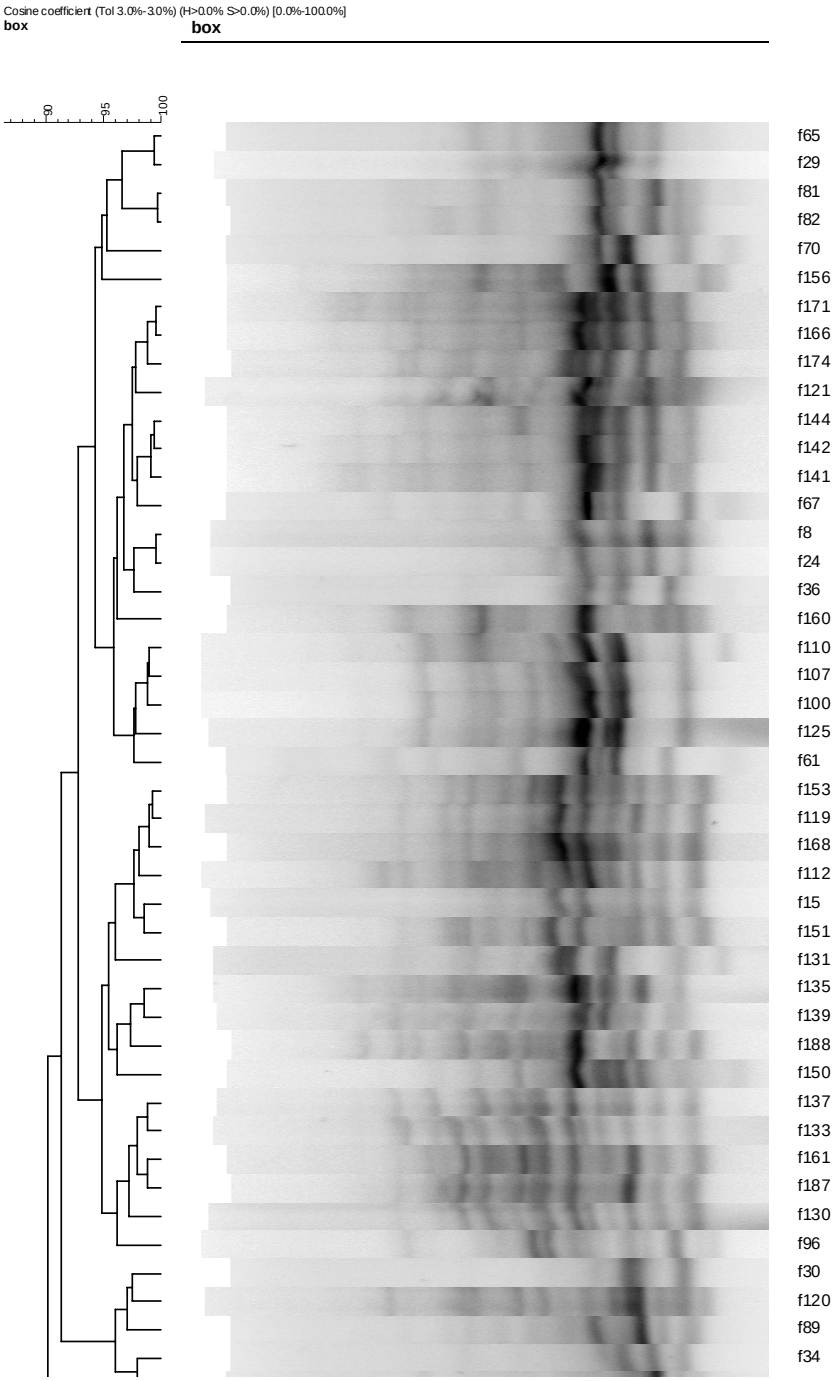

Cosine coefficient (Tol 3.0%-3.0%) (H>0.0% S>0.0%) [0.0%-100.0%]
box
box
90
95
100
f65
f29
f81
f82
f70
f156
f171
f166
f174
f121
f144
f142
f141
f67
f8
f24
f36
f160
f110
f107
f100
f125
f61
f153
f119
f168
f112
f15
f151
f131
f135
f139
f188
f150
f137
f133
f161
f187
f130
f96
f30
f120
f89
f34

Figure A.4 (continued)

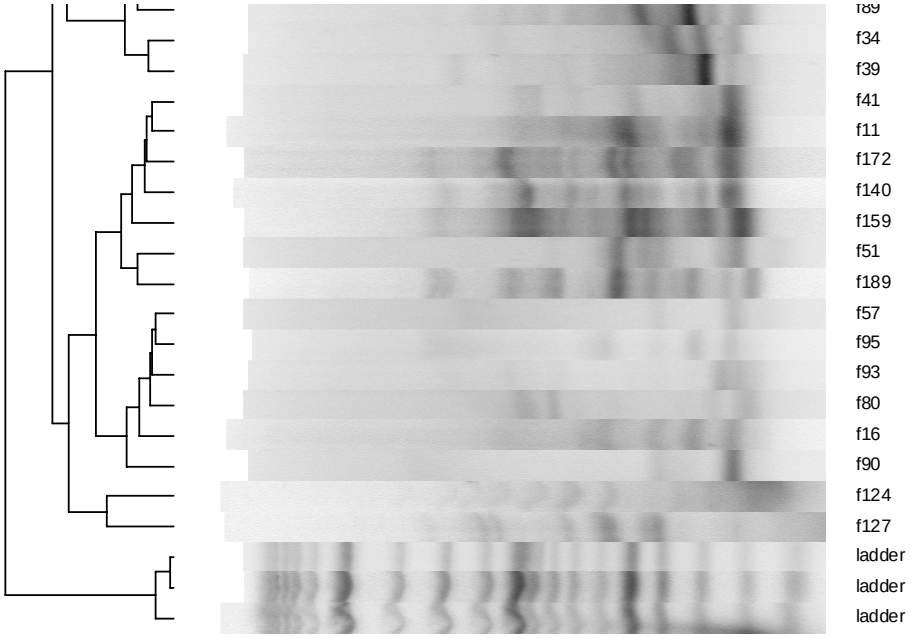


Table A.4. Classification of isolates from Ft. Pierce, FL based on REP-PCR patterns and *rpoB* sequence data. [a] The isolate number for this soil sample. Throughout Chapter 5 these isolates will be proceeded with the letter "f". [b] The OTU to which they belong, as determined by the program DOTUR. [c] REP indicates screening via REP-PCR, and those chosen for sequencing at *rpoB* for OTU determination are indicated with "seq".

Isolate[a]	OTU[b]	Screen[c]
8	OTU 24	REP
11	OTU 25	REP, seq
15	OTU 26	REP, seq
16	OTU 26	REP
24	OTU 24	REP
29	OTU 27	REP, seq
30	OTU 24	REP, seq
34	OTU 24	REP, seq
36	OTU 24	REP
39	OTU 14	REP, seq
41	OTU 25	REP, seq
51	OTU 1	REP, seq
57	OTU 24	REP, seq
61	OTU 1	REP, seq
65	OTU 27	REP
67	OTU 1	REP, seq
70	OTU 24	REP, seq
80	OTU 26	REP, seq
81	OTU 25	REP, seq
82	OTU 25	REP
89	OTU 26	REP, seq
90	OTU 24	REP, seq
93	OTU 21	REP, seq
95	OTU 24	REP, seq
96	OTU 28	REP, seq
100	OTU 24	REP, seq
107	OTU 24	REP
110	OTU 24	REP
112	OTU 26	REP
119	OTU 26	REP
120	OTU 24	REP, seq
121	OTU 24	REP
124	OTU 26	REP
125	OTU 24	REP
127	OTU 24	REP, seq
130	OTU 26	REP
131	OTU 24	REP
133	OTU 26	REP
135	OTU 24	REP, seq
137	OTU 26	REP
139	OTU 24	REP
140	OTU 24	REP
141	OTU 24	REP, seq
142	OTU 24	REP
144	OTU 24	REP
150	OTU 1	REP, seq
151	OTU 26	REP
153	OTU 26	REP
156	OTU 3	REP, seq
159	OTU 25	REP, seq
160	OTU 25	REP, seq
161	OTU 26	REP
166	OTU 24	REP
168	OTU 26	REP
171	OTU 24	REP
174	OTU 24	REP
187	OTU 26	REP
188	OTU 26	REP
189	OTU 25	REP, seq
192	OTU 24	seq
193	OTU 24	seq
194	OTU 24	seq
198	OTU 24	seq
201	OTU 24	seq
202	OTU 24	seq
203	OTU 24	seq
204	OTU 24	seq
208	OTU 24	seq
209	OTU 24	seq
215	OTU 24	seq
219	OTU 24	seq
220	OTU 24	seq
221	OTU 24	seq
222	OTU 24	seq
224	OTU 24	seq
225	OTU 24	seq
231	OTU 24	seq
232	OTU 24	seq
233	OTU 24	seq
234	OTU 24	seq
235	OTU 24	seq
236	OTU 24	seq
237	OTU 24	seq
238	OTU 24	seq
239	OTU 24	seq
240	OTU 24	seq
243	OTU 24	seq
244	OTU 24	seq
245	OTU 24	seq
248	OTU 40	seq
251	OTU 24	seq
252	OTU 27	seq
253	OTU 24	seq
255	OTU 24	seq
257	OTU 24	seq

Figure A.5. Dendrogram of REP-PCR patterns for streptomycetes isolated from Astoria, OR. All isolates shown are from Astoria, Oregon, whether or not proceeded by “OR”.

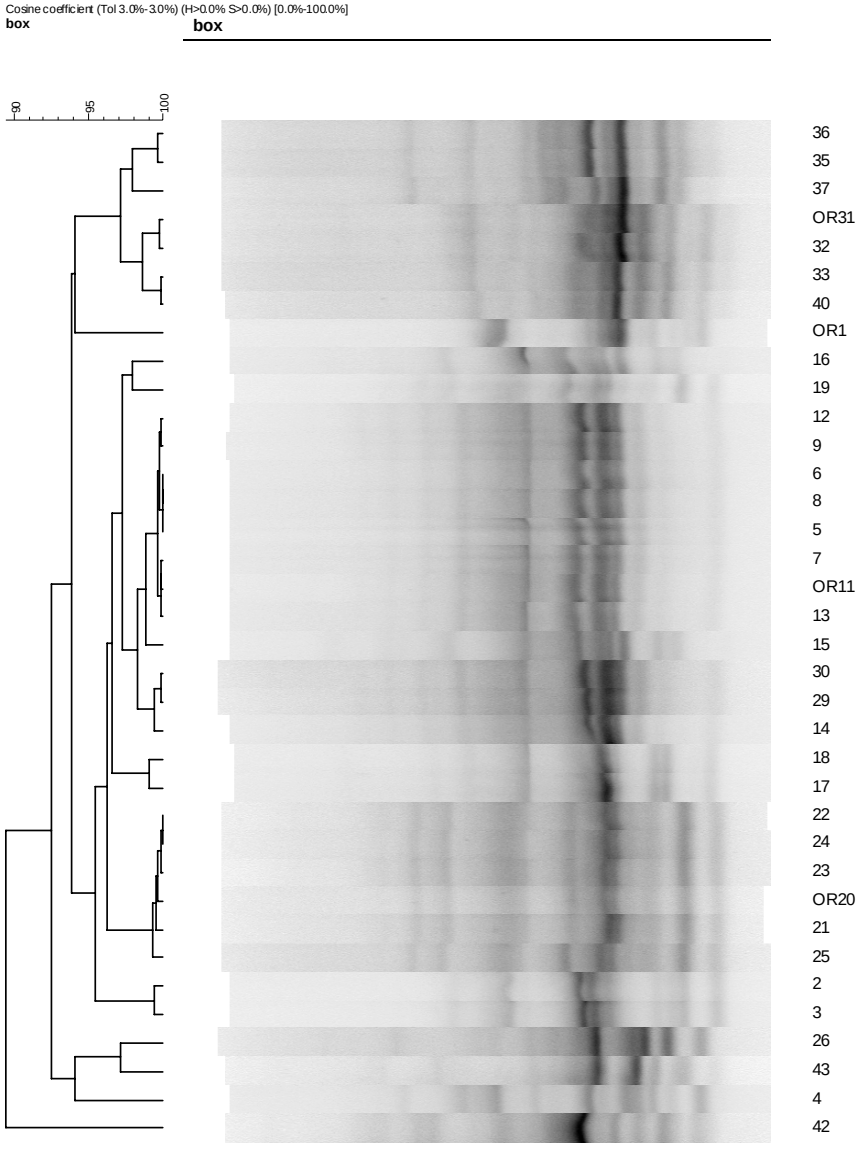

Cosine coefficient (Tol 3.0%-3.0%) (H>0.0% S>0.0%) [0.0%-100.0%]
box
box
90
95
100
36
35
37
OR31
32
33
40
OR1
16
19
12
9
6
8
5
7
OR11
13
15
30
29
14
18
17
22
24
23
OR20
21
25
2
3
26
43
4
42

Table A.5. Classification of isolates from Astoria, OR based on REP-PCR patterns and *rpoB* sequence data. [a] The isolate number for this soil sample. Throughout Chapter 5 these isolates will be proceeded with the letters "or". [b] The OTU to which they belong, as determined by the program DOTUR. [c] REP indicates screening via REP-PCR, and those chosen for sequencing at rpoB for OTU determination are indicated with "seq".

Isolate[a]	OTU[b]	Screen[c]
1	OTU 23	REP, seq
2	OTU 15	REP, seq
3	OTU 15	REP
4	OTU 32	REP, seq
5	OTU 5	REP
6	OTU 5	REP
7	OTU 5	REP
8	OTU 5	REP
9	OTU 5	REP, seq
11	OTU 5	REP
12	OTU 5	REP
13	OTU 5	REP, seq
14	OTU 5	REP
15	OTU 4	REP
16	OTU 15	REP
17	OTU 15	REP, seq
18	OTU 15	REP
19	OTU 15	REP, seq
20	OTU 15	REP, seq
21	OTU 15	REP, seq
22	OTU 15	REP
23	OTU 15	REP
24	OTU 15	REP
25	OTU 15	REP
26	OTU 31	REP, seq
29	OTU 5	REP, seq
30	OTU 5	REP, seq
31	OTU 34	REP, seq
32	OTU 34	REP, seq
33	OTU 34	REP, seq
35	OTU 34	REP
36	OTU 34	REP
37	OTU 34	REP
40	OTU 34	REP, seq
42	OTU 35	REP
43	OTU 33	REP
48	OTU 77	seq
49	OTU 74	seq
50	OTU 74	seq
51	OTU 75	seq
52	OTU 74	seq
53	OTU 74	seq
54	OTU 78	seq
55	OTU 79	seq
56	OTU 75	seq
57	OTU 74	seq
58	OTU 33	seq
59	OTU 33	seq
60	OTU 74	seq
61	OTU 80	seq
62	OTU 74	seq
63	OTU 34	seq
64	OTU 34	seq
66	OTU 81	seq
68	OTU 74	seq
69	OTU 74	seq
70	OTU 74	seq
73	OTU 74	seq
74	OTU 77	seq
75	OTU 80	seq
76	OTU 74	seq
77	OTU 80	seq
78	OTU 74	seq
80	OTU 77	seq
82	OTU 76	seq
83	OTU 74	seq
84	OTU 76	seq
85	OTU 74	seq
86	OTU 15	seq
87	OTU 74	seq
88	OTU 33	seq
89	OTU 80	seq
90	OTU 79	seq
91	OTU 74	seq
92	OTU 74	seq
93	OTU 74	seq
94	OTU 74	seq
95	OTU 74	seq
96	OTU 77	seq
101	OTU 74	seq
102	OTU 74	seq
103	OTU 75	seq
104	OTU 75	seq
105	OTU 56	seq
106	OTU 34	seq
107	OTU 74	seq
108	OTU 74	seq
109	OTU 56	seq
110	OTU 75	seq
111	OTU 74	seq
112	OTU 74	seq
113	OTU 74	seq
114	OTU 76	seq
115	OTU 75	seq
117	OTU 34	seq

Figure A.6. Dendrogram of REP-PCR patterns for streptomycetes isolated from Caldwell Field, Ithaca, NY. All isolates shown, except those labeled as “ladder”, are from Caldwell Field, Ithaca, NY.

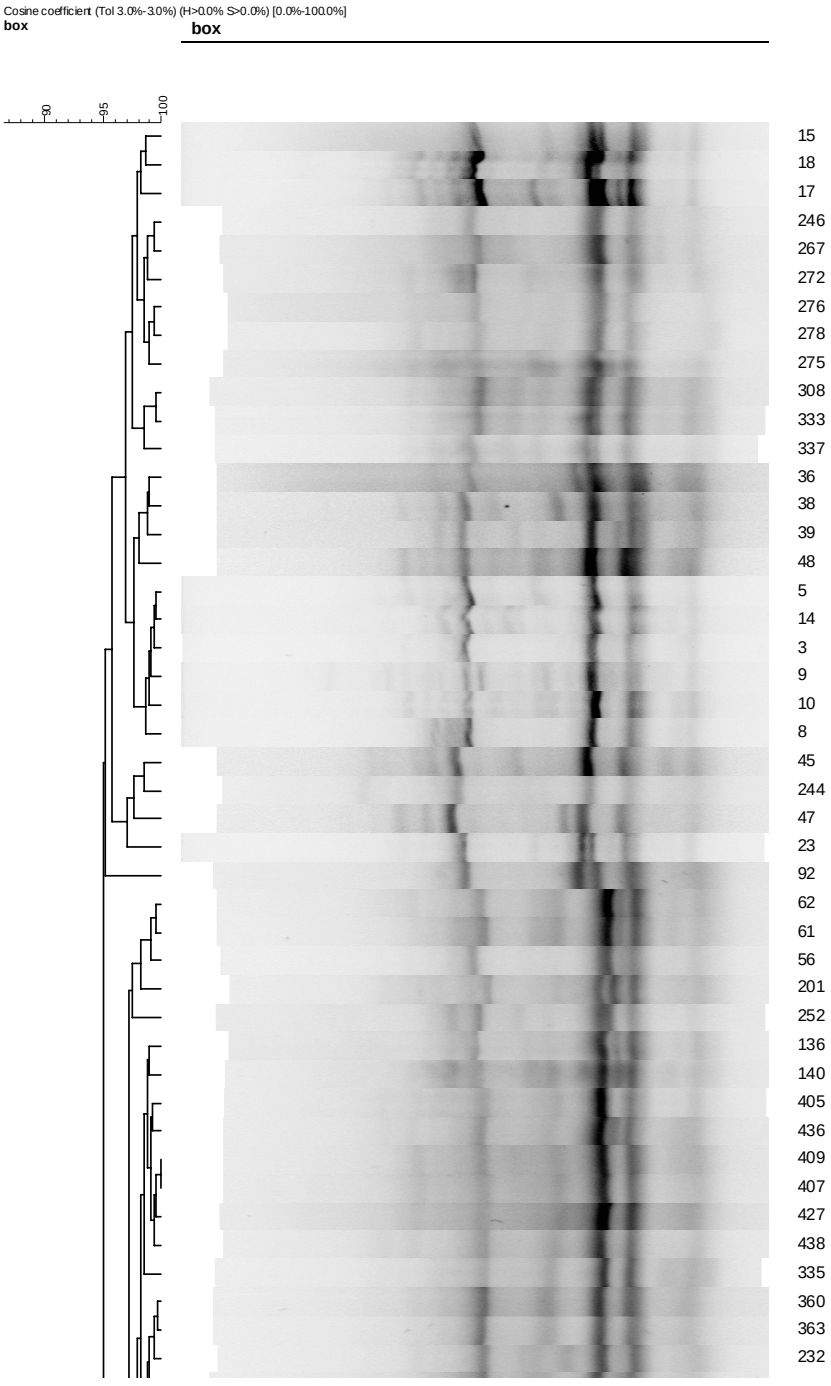

Cosine coefficient (Tol 3.0%-3.0%) (H>0.0% S>0.0%) [0.0%-100.0%]
box
box
90
95
100
15
18
17
246
267
272
276
278
275
308
333
337
36
38
39
48
5
14
3
9
10
8
45
244
47
23
92
62
61
56
201
252
136
140
405
436
409
407
427
438
335
360
363
232

Figure A.6 (continued)

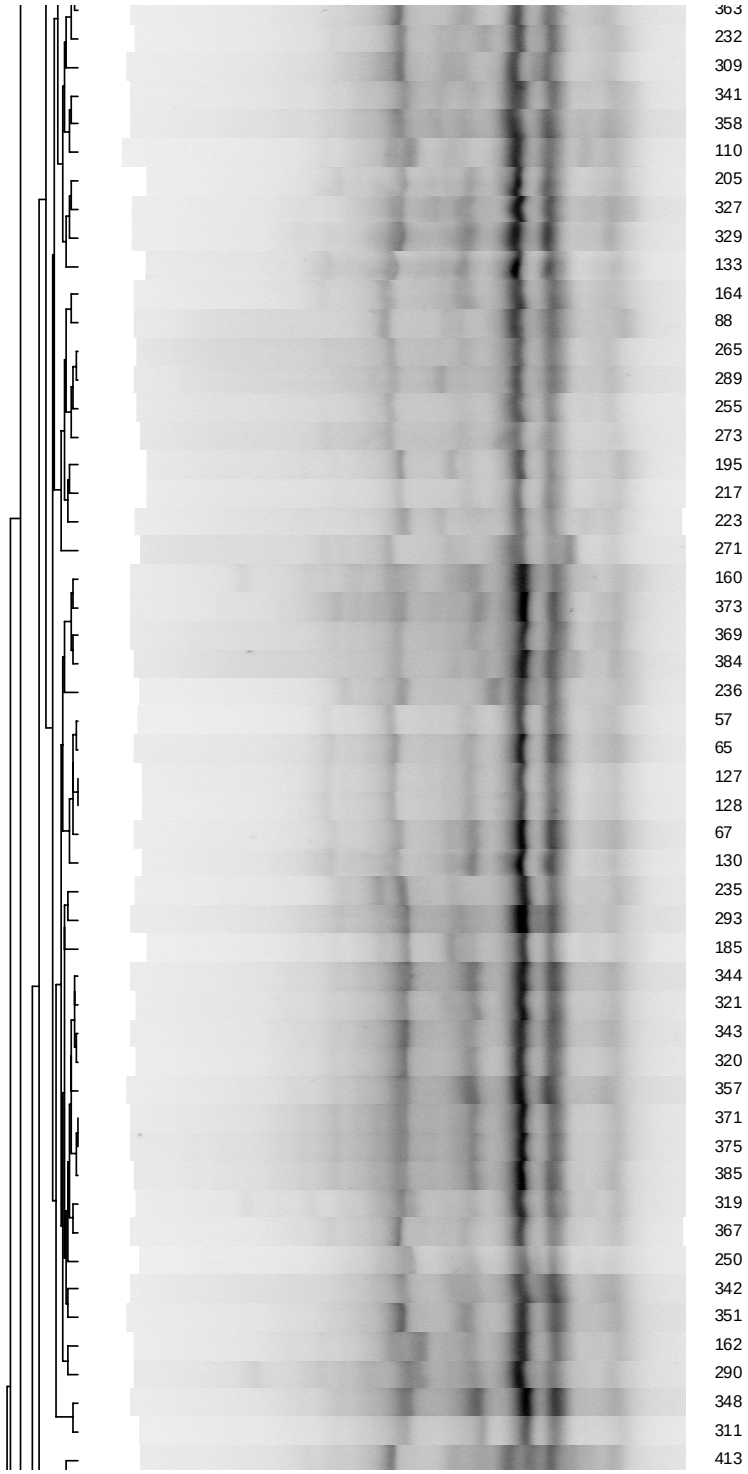


Figure A.6 (continued)

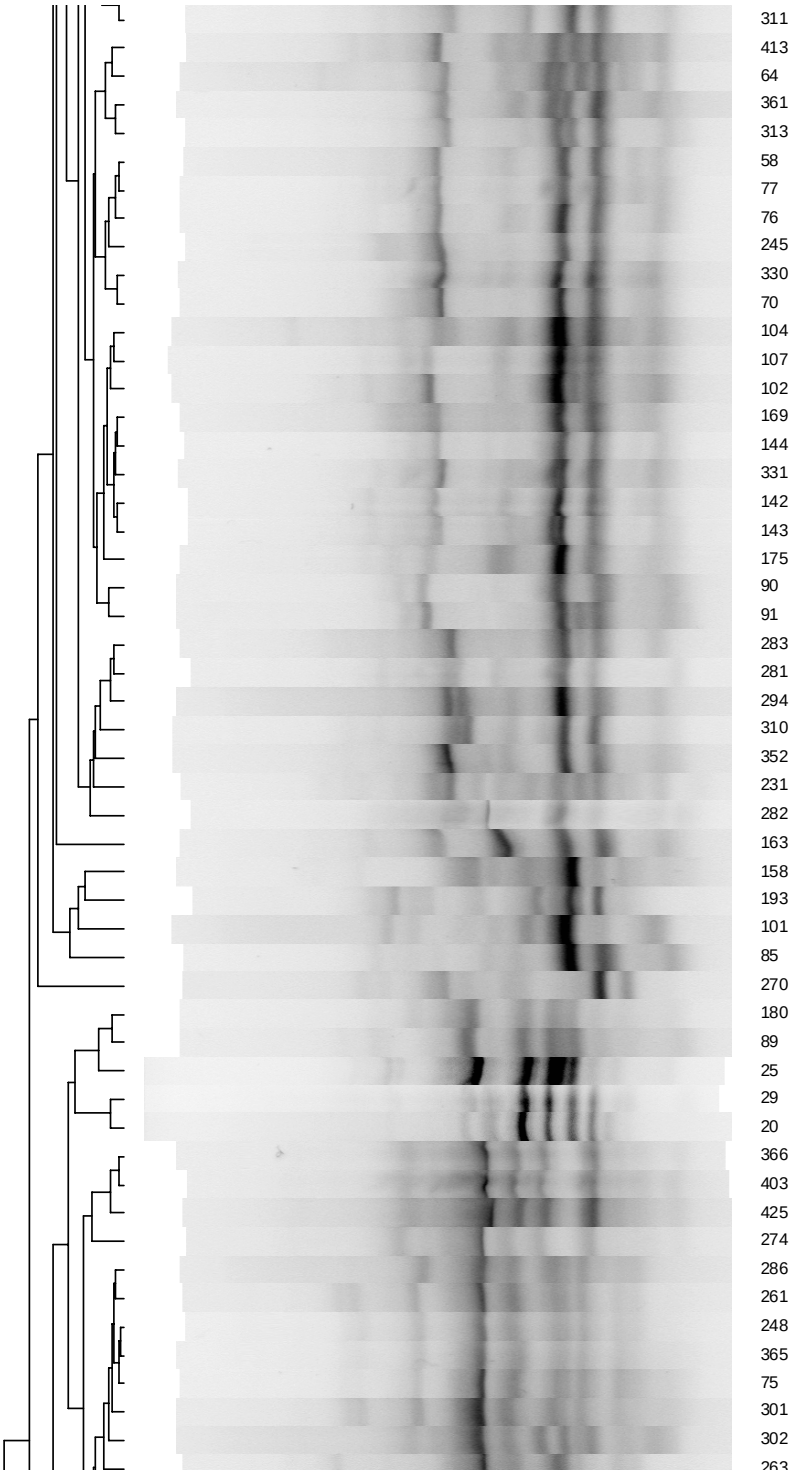


Figure A.6 (continued)

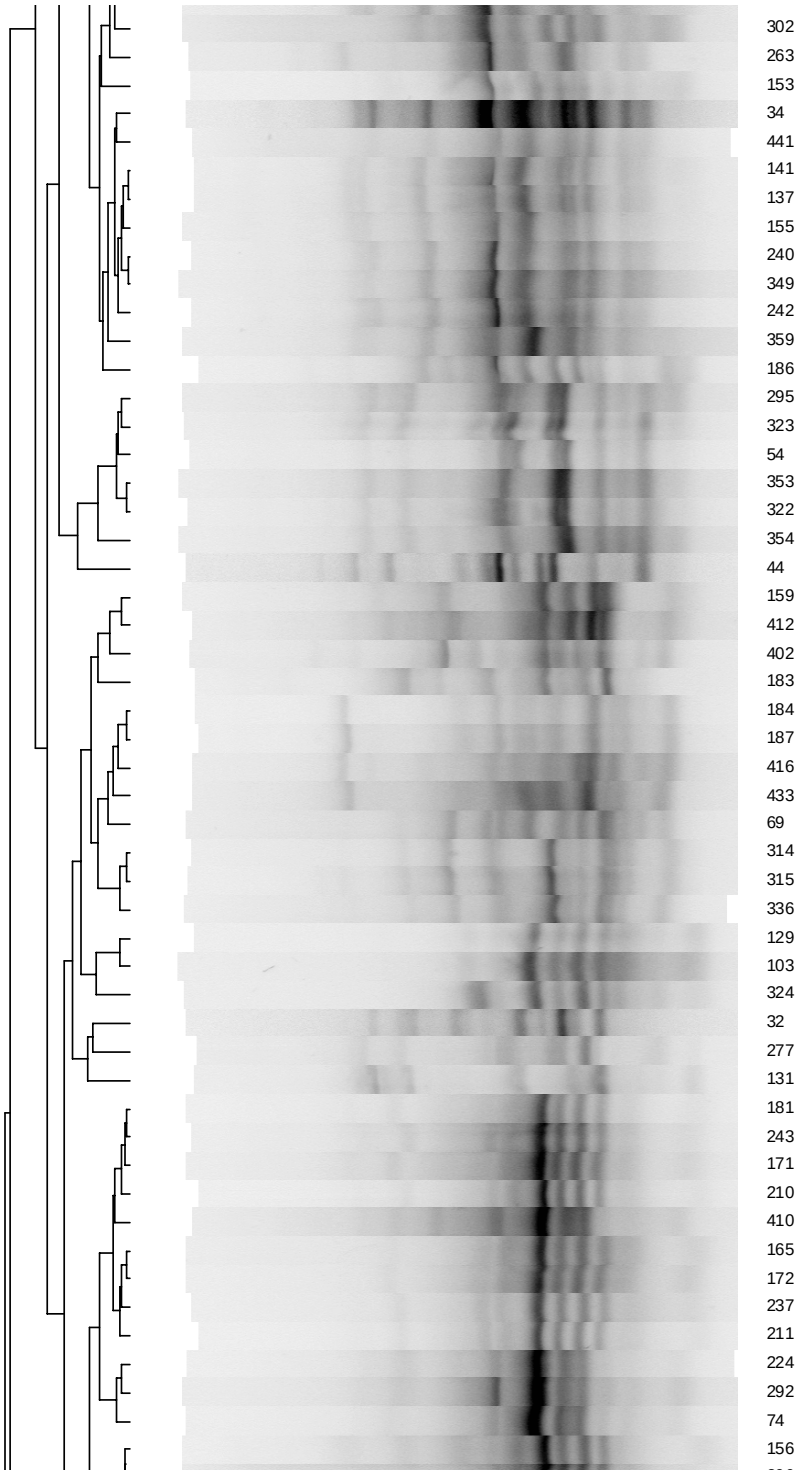


Figure A.6 (continued)

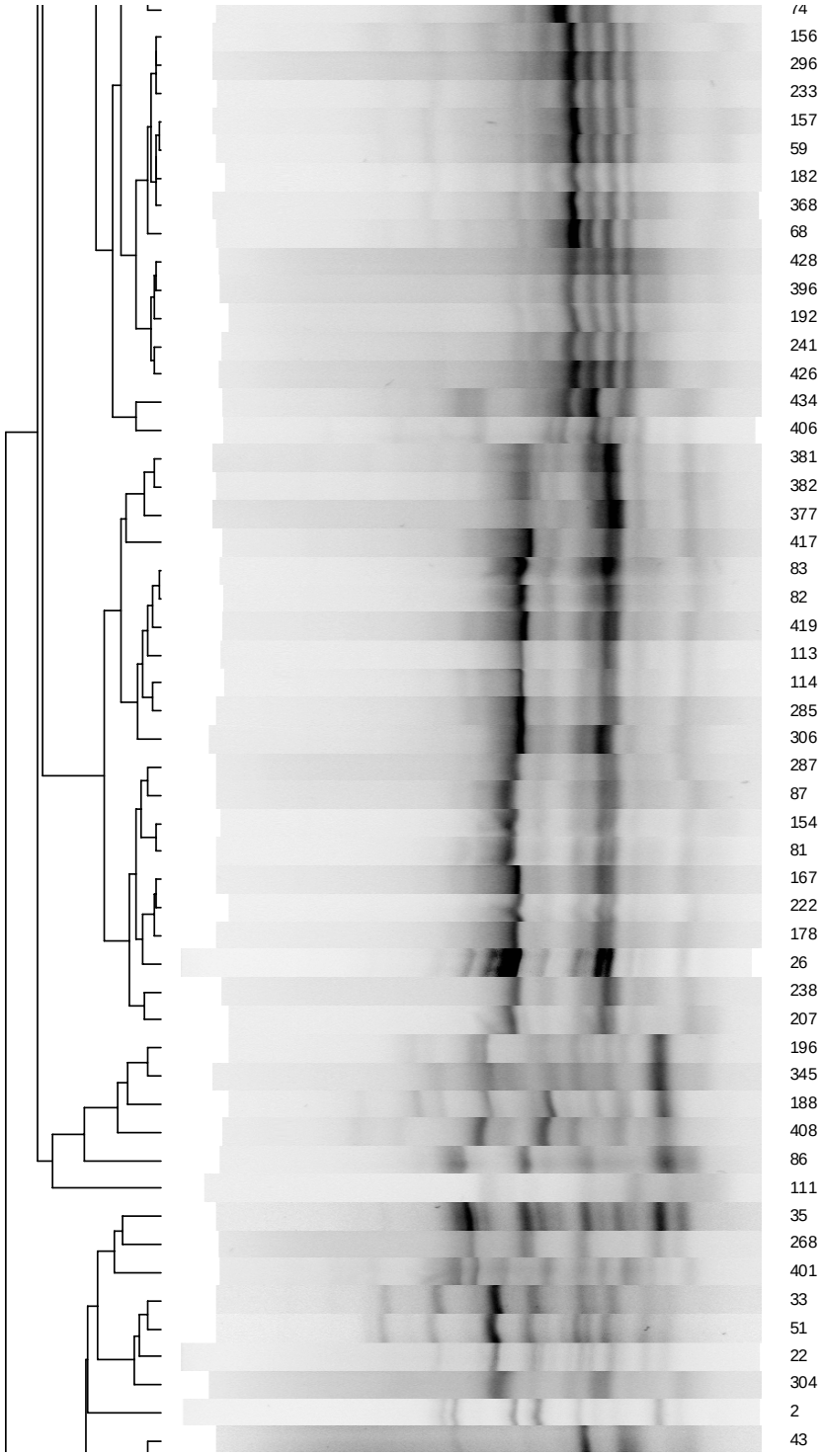


Figure A.6 (continued)

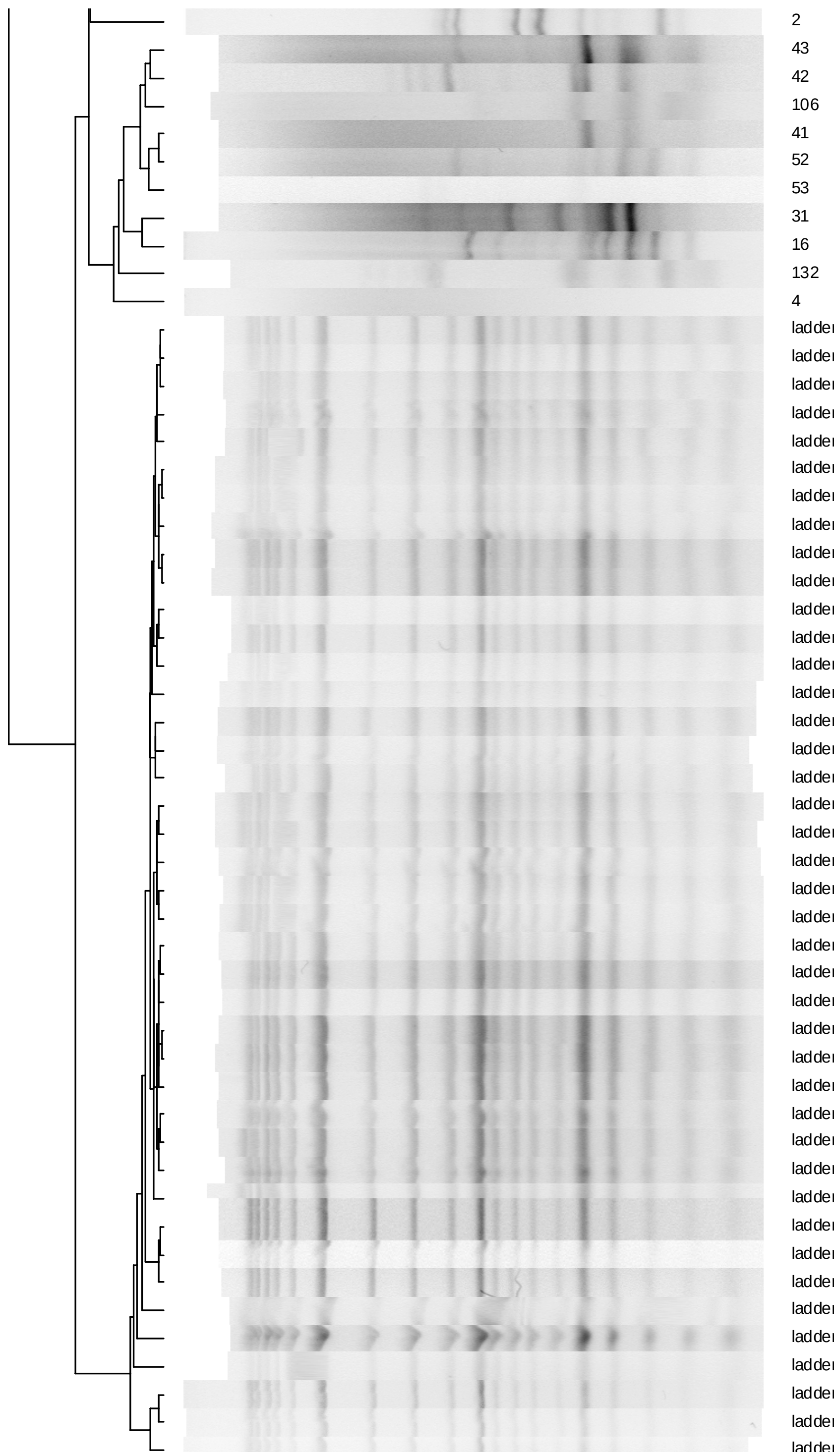


Table A.6. Classification of isolates from Caldwell Field, Ithaca, NY based on REP-PCR patterns, colony morphology and *rpoB* sequence data. [a] The isolate number for this soil sample. Caldwell Field isolates are referred to only with the isolate number throughout Chapter 5. [b] The OTU to which they belong, as determined by the program DOTUR. [c] "REP" indicates screening via REP-PCR, "morph" indicates screening by colony morphology, and those chosen for sequencing at rpoB for OTU determination are indicated with "seq".

Isolate[a]	OTU[b]	Screen[c]
2	OTU 11	REP, seq
3	OTU 3	REP, seq
4	OTU 3	REP, seq
5	OTU 3	REP, seq
8	OTU 3	REP
9	OTU 3	REP, seq
10	OTU 3	REP, seq
14	OTU 3	REP
15	OTU 3	REP, seq
16	OTU 6	REP, seq
17	OTU 3	REP, seq
18	OTU 3	REP
20	OTU 5	REP, seq
22	OTU 4	REP, seq
23	OTU 3	REP
25	OTU 5	REP, seq
26	OTU 1	REP, seq
29	OTU 5	REP, seq
31	OTU 2	REP, seq
32	OTU 2	REP, seq
33	OTU 4	REP, seq
34	OTU 4	REP, seq
35	OTU 11	REP, seq
36	OTU 3	REP, seq
38	OTU 3	REP, seq
39	OTU 3	REP, seq
41	OTU 3	REP, seq
42	OTU 3	REP, seq
43	OTU 3	REP, seq
44	OTU 12	REP, seq
45	OTU 3	REP, seq
47	OTU 3	REP, seq
48	OTU 3	REP, seq
51	OTU 4	REP, seq
52	OTU 6	REP, seq
53	OTU 3	REP, seq
54	OTU 12	REP, seq
56	OTU 3	REP, seq
57	OTU 3	REP
58	OTU 3	REP, seq
59	OTU 5	REP, seq
61	OTU 3	REP, seq
62	OTU 3	REP, seq
64	OTU 6	REP, seq
65	OTU 3	REP
67	OTU 3	REP, seq
68	OTU 5	REP, seq
69	OTU 6	REP, seq
70	OTU 3	REP, seq
74	OTU 5	REP, seq
75	OTU 4	REP
76	OTU 3	REP
77	OTU 3	REP
81	OTU 1	REP, seq
82	OTU 1	REP, seq
83	OTU 1	REP, seq
85	OTU 3	REP, seq
86	OTU 11	REP, seq
87	OTU 1	REP, seq
88	OTU 3	REP
89	OTU 5	REP
90	OTU 3	REP
91	OTU 3	REP, seq
92	OTU 3	REP, seq
102	OTU 3	REP
103	OTU 5	REP
104	OTU 3	REP
106	OTU 3	REP
107	OTU 3	REP
110	OTU 3	REP, seq
113	OTU 1	REP, seq
114	OTU 1	REP
127	OTU 3	REP
128	OTU 3	REP
129	OTU 5	REP
130	OTU 3	REP
131	OTU 2	REP, seq
132	OTU 3	REP, seq
133	OTU 3	REP
136	OTU 3	REP, seq
137	OTU 4	REP, seq
140	OTU 3	REP
141	OTU 4	REP, seq
142	OTU 3	REP, seq
143	OTU 3	REP, seq
144	OTU 3	REP, seq
153	OTU 5	REP, seq
154	OTU 1	REP, seq
155	OTU 4	REP, seq
156	OTU 5	REP, seq

Table A.6 (continued)

Isolate[a]	OTU[b]	Screen[c]	Isolate[a]	OTU[b]	Screen[c]
157	OTU 5	REP, seq	243	OTU 5	REP
159	OTU 3	REP, seq	244	OTU 3	REP, seq
160	OTU 3	REP	245	OTU 3	REP, seq
162	OTU 3	REP, seq	246	OTU 3	REP, seq
163	OTU 3	REP, seq	248	OTU 4	REP, seq
164	OTU 3	REP, seq	250	OTU 3	REP
165	OTU 5	REP, seq	252	OTU 3	REP
167	OTU 1	REP, seq	255	OTU 3	REP
169	OTU 3	REP, seq	261	OTU 4	REP, seq
171	OTU 5	REP, seq	263	OTU 4	REP, seq
172	OTU 5	REP, seq	265	OTU 3	REP, seq
175	OTU 3	REP	267	OTU 3	REP
178	OTU 1	REP, seq	268	OTU 11	REP, seq
180	OTU 5	REP, seq	270	OTU 10	REP, seq
181	OTU 5	REP, seq	271	OTU 3	REP
182	OTU 5	REP	272	OTU 3	REP
183	OTU 2	REP, seq	273	OTU 3	REP
184	OTU 7	REP, seq	274	OTU 8	REP, seq
185	OTU 3	REP	275	OTU 3	REP
186	OTU 8	REP, seq	276	OTU 3	REP
187	OTU 7	REP, seq	277	OTU 12	REP, seq
188	OTU 9	REP, seq	278	OTU 3	REP
192	OTU 5	REP	281	OTU 3	REP
193	OTU 2	REP, seq	282	OTU 3	REP, seq
195	OTU 3	REP, seq	283	OTU 3	REP
196	OTU 10	REP, seq	285	OTU 1	REP
201	OTU 3	REP, seq	286	OTU 4	REP, seq
205	OTU 3	REP, seq	287	OTU 1	REP, seq
207	OTU 1	REP, seq	289	OTU 3	REP
210	OTU 5	REP	290	OTU 3	REP, seq
211	OTU 5	REP	292	OTU 5	REP, seq
217	OTU 3	REP	293	OTU 3	REP
222	OTU 1	REP	294	OTU 3	REP, seq
223	OTU 3	REP	295	OTU 12	REP, seq
224	OTU 5	REP, seq	296	OTU 5	REP
231	OTU 6	REP, seq	301	OTU 4	REP, seq
232	OTU 3	REP, seq	302	OTU 8	REP, seq
233	OTU 5	REP	304	OTU 1	REP, seq
235	OTU 3	REP	306	OTU 1	REP
236	OTU 3	REP, seq	308	OTU 3	REP
237	OTU 5	REP	309	OTU 3	REP
238	OTU 1	REP	310	OTU 3	REP
240	OTU 4	REP	311	OTU 3	REP
241	OTU 5	REP	313	OTU 3	REP, seq
242	OTU 4	REP, seq	314	OTU 6	REP, seq

Table A.6 (continued)

Isolate[a]	OTU[b]	Screen[c]	Isolate[a]	OTU[b]	Screen[c]
315	OTU 6	REP, seq	396	OTU 5	REP
316	OTU 3	REP	401	OTU 6	REP, seq
319	OTU 3	REP, seq	402	OTU 6	REP, seq
320	OTU 3	REP	403	OTU 8	REP, seq
321	OTU 3	REP	405	OTU 3	REP
322	OTU 12	REP, seq	407	OTU 3	REP
323	OTU 12	REP, seq	408	OTU 9	REP, seq
327	OTU 3	REP	409	OTU 3	REP
329	OTU 3	REP, seq	410	OTU 5	REP, seq
330	OTU 3	REP, seq	412	OTU 3	REP
331	OTU 3	REP	413	OTU 6	REP
333	OTU 3	REP	416	OTU 7	REP, seq
335	OTU 3	REP	417	OTU 1	REP
336	OTU 6	REP, seq	419	OTU 1	REP
337	OTU 3	REP	425	OTU 8	REP, seq
341	OTU 3	REP	426	OTU 5	REP
342	OTU 3	REP	427	OTU 3	REP
343	OTU 3	REP	428	OTU 5	REP
344	OTU 3	REP	433	OTU 7	REP, seq
345	OTU 10	REP, seq	434	OTU 3	REP
348	OTU 3	REP	436	OTU 3	REP
349	OTU 4	REP	438	OTU 3	REP
351	OTU 3	REP	441	OTU 4	REP
352	OTU 3	REP	442	OTU 5	REP
353	OTU 12	REP, seq	443	OTU 3	morph
354	OTU 12	REP, seq	446	OTU 3	morph
357	OTU 3	REP	447	OTU 6	REP, seq
358	OTU 3	REP	448	OTU 3	morph
359	OTU 4	REP	449	OTU 12	REP, seq
360	OTU 3	REP	450	OTU 3	morph
361	OTU 3	REP	452	OTU 12	REP, seq
363	OTU 3	REP	453	OTU 3	morph
365	OTU 4	REP	454	OTU 8	REP, seq
366	OTU 8	REP, seq	455	OTU 3	morph
367	OTU 3	REP	456	OTU 6	REP, seq
368	OTU 5	REP	457	OTU 3	morph
369	OTU 3	REP	458	OTU 12	REP, seq
371	OTU 3	REP	459	OTU 6	REP, seq
373	OTU 3	REP	460	OTU 3	morph
375	OTU 3	REP	461	OTU 3	morph
377	OTU 1	REP	465	OTU 13	REP, seq
381	OTU 1	REP	466	OTU 6	REP, seq
382	OTU 1	REP	467	OTU 3	morph
384	OTU 3	REP	468	OTU 3	morph
385	OTU 3	REP	469	OTU 3	morph

Table A.6 (continued)

Isolate[a]	OTU[b]	Screen[c]
470	OTU 3	morph
471	OTU 3	morph
472	OTU 3	morph
473	OTU 3	morph
474	OTU 3	morph
478	OTU 3	morph
479	OTU 3	morph
480	OTU 5	REP
481	OTU 7	REP, seq
482	OTU 3	morph
484	OTU 3	morph
485	OTU 3	morph
487	OTU 3	morph
489	OTU 3	morph
490	OTU 3	morph
491	OTU 3	morph
492	OTU 3	morph
493	OTU 3	morph
494	OTU 3	morph
495	OTU 3	morph
496	OTU 3	morph
502	OTU 1	REP
503	OTU 1	REP
505	OTU 3	morph
506	OTU 4	REP
507	OTU 3	morph
508	OTU 12	REP, seq
509	OTU 3	morph
512	OTU 3	morph
513	OTU 3	morph
514	OTU 3	morph
515	OTU 1	REP
516	OTU 4	REP
518	OTU 1	REP
519	OTU 5	REP, seq
520	OTU 3	morph
522	OTU 3	morph
524	OTU 1	REP
526	OTU 3	morph
527	OTU 3	morph
529	OTU 4	REP
530	OTU 3	morph
534	OTU 1	REP
536	OTU 4	REP
537	OTU 3	morph
538	OTU 3	morph
540	OTU 1	REP
544	OTU 12	REP, seq
545	OTU 3	morph
546	OTU 6	REP
547	OTU 3	morph
548	OTU 3	morph
549	OTU 6	REP
554	OTU 3	morph
556	OTU 3	morph
557	OTU 3	morph
558	OTU 6	REP
559	OTU 3	morph
561	OTU 10	REP, seq
563	OTU 4	REP
564	OTU 5	REP
567	OTU 5	REP
568	OTU 5	REP

Table A.7. The list of all isolates included in the 13 site *rpoB* data set, divided into OTUs by the program DOTUR with a furthest neighbor clustering cutoff of 0.01.

OTU 1	113, 154, 167, 178, 207, 26, 287, 304, 324, 81, 83, 87, 579, 611, 629
OTU 2	131, 183, 193, 31, 32, b12b, b49, b94
OTU 3	132, 15, 45, 53, ch27, ch34, char2, sun61, ms142, 163, 201, 265, 294, 43, 47, 5, 62, 67, ch14, ch20, ms107, m55, 245, 290, 9, 143, 159, 17, 39, f156, sun64, 169, 232, 236, 244, 246, 313, 319, 329, 330, 38, 41, 42, 70, 85, 91, 92, b81, ch13, ch21, ch33, ch38, ch39, ch40, t109, cald738, sun89, m38, 282, ms194
OTU 4	137, 155, 22, 242, 248, 261, 263, 286, 301, 33, 34, 51, or15, 122man, ms137, b202
OTU 5	153, 180, 224, 25, 292, 410, 74, b11, b126, b203, or13, or9, st123, st153, st170, st177, t1, t99, or29, or30, sun10, sun13, sun9, ms163, ms172, ms175, ms180, 519, 156, b62, ms168, 157, 165, 171, 172, 20, 29, 59, 68, st133
OTU 6	16, 231, 401, 402, 447, 456, 459, 466, 52, 64, 69, 314, 315, 336
OTU 7	184, 187, 416, 433, 481, b58, st184, sun20, sun49, sun54, ms100, ms102, ms103, ms104, ms105, ms108, ms109, ms116, ms119, ms120, ms122, ms126, ms128, ms129, ms130, ms144, ms148, ms149, ms153, ms159, ms169, ms171, ms177, ms182, ms206, ms51, ms52, ms54, ms81, ms84, ms91, ms97, 585, 624, b117
OTU 8	186, 302, ms135, ms93, 274, 366, 403, 425, 454, 570
OTU 9	188, 408
OTU 10	196, 561, 270, 345, t26, t54, t63, t39
OTU 11	2, 268, 35, 86, b5
OTU 12	277, 322, 323, 353, 354, 458, 54, 544, gb2, man155, cald772, sun110, sun2, cald784, 44, 449, 508, gb12, 295, 452, sun112, sun71
OTU 13	465, st92
OTU 14	b150, b76, st180, sun137, sun8, sun88, b63, st139, ms101, ms146, ms71, ms88, f39, ms114
OTU 15	b214, b71, or3, b84, b90, or2, or86, b89, or17, man185, or19, or20, or21
OTU 16	b46
OTU 17	b47
OTU 18	ch17, m75
OTU 19	ch19, 184man
OTU 20	ch23, w49, w51, gb1, m21, uw11, 11man, 145man, 154man, 162man, 164man, 165man, 166man, 173man, 179man, 180man, 188man, 189man, 196man, 205man, 208man, 209man, 24man, 26man, 27man, 30man, 69man, 70man, man73, man75, man78, man7, uw13, m58, uw12, uw27
OTU 21	ch26, ch28, ch31, st210, sun101, sun114, sun115, sun116, sun117, sun122, sun136, sun140, sun81, ms196, f93, ch29
OTU 22	ch37, gb3, ms198, ms145, ms195, m107, m108, m115, m117, m118, m120, m121, m124
OTU 23	ch4, st172, st201, or1
OTU 24	f100, f127, f57, f70, f90, st72, t101, t14, t40, t52, ms155, f203, f233, f240, f30, f95, f8, f219, f234, f232, ms164, t66, t66.dup, f120, t49, f253, f135, f141, f34, f174, f24, f192, f193, f194, f198, f201, f202, f204, f208, f209, f215, f220, f221, f222, f224, f225, f231, f235, f236, f237, f238, f239, f243, f244, f245, f251, f255, f257
OTU 25	f11, f159, uw102, uw104, uw39, ms187, ms189, ms199, ms200, ms201, ms208, ms209, ms210, f41, ms202, uw2, uw21, f160, uw103, uw7, sun103, sun93, ms30, ms53, sun142, f189, f81, f121
OTU 26	f15, f80, f89, st146, t11, t32, t56, t78, f130, t20

Table A.7 (continued)

OTU 27	f150, st140, st196, f51, gb15, ms152, f61, f67, gb14, ms115, ms181, ms98, f252, ms140, ms162, ms184, ms183, st115, st205
OTU 28	f29, ms205
OTU 29	f96
OTU 30	gb10, gb11, sun100, sun104, sun105, sun109, sun127, sun73, sun74, sun80, sun87
OTU 31	or26
OTU 32	or4
OTU 33	or43, or58, or59, or88
OTU 34	or40, or33, or106, or117, or63, or64, m112, m119, m2, m46, m64, m65, m71, m81, m87, m9, m91, w27, w35, w4, or31, or32
OTU 35	or42, w11, w16, w17, w26, w28, w29, w3, w39, w5, w54, w59, w61, w63, w12, w14, w6
OTU 36	st125, sun94, st70
OTU 37	st134, st96, t12, t127, t18, t34, sun113, sun125, sun129, sun138, sun143, sun56, sun63, sun66, sun86, sun118
OTU 38	st150, st77, st82
OTU 39	st151
OTU 40	st165, st81, t73, t22, f248
OTU 41	st173
OTU 42	st207, st85
OTU 43	st211
OTU 44	st212
OTU 45	st213
OTU 46	t120, t96, t15, t155, t178, t30
OTU 47	t128, t35
OTU 48	t168, t51
OTU 49	t3, t3.dup
OTU 50	t43
OTU 51	t55, man176, man92
OTU 52	t79
OTU 53	t89
OTU 54	t92
OTU 55	uw100, uw24, uw30, uw70, uw67, uw72
OTU 56	uw60, or105, or109, m11, m111, m123, m31, m34, m4, m40, m5, m56, m68, m7, m72, m76, m96, w24, w55, w7, m44, m82
OTU 57	uw82
OTU 58	105man, 170man, 147man, 167man, 171man, 181man, 201man, 202man, 203man, 6man
OTU 59	10man, 116man, 118man, 121man, 125man, 134man, 136man, 137man, 138man, 139man, 13man, 141man, 142man, 143man, 144man, 150man, 152man, 153man, 156man, 158man, 163man, 186man, 194man, 1man, 25man, 32man, 34man, 35man, 36man, 3man, 48man, 4man, 55man, 56man, 58man, 5man, 61man, 62man, 72man, 83man, 84man, 85man, den44, den46, den48, den1, den10, den11, den12, den13, den14, den15, den16, den17, den19, den20, den22, den23, den24, den26, den27, den28, den29, den3, den30, den31, den32, den33, den34, den35, den36, den37, den38, den39, den4, den41, den42, den5, den6, den7, den8, den9, w18, w19, w21, w34, w44, w46, w9, sun121, sun134, sun75, 161man, 110man, 119man, 21man, 37man, 71man, 82man, 93man, w30, w31, w32, w60, 19man

Table A.7 (continued)

OTU 60	172man, 195man, m109, m25, m37, m43, m47, m53, m57, m60, m66, m67, m77, m85, m90, m93, m95
OTU 61	sun107, sun97, w25, w41, w42
OTU 62	sun130, ms192, sun51, ms154, ms157
OTU 63	sun15, sun45
OTU 64	sun58, ms112, ms123, ms143, ms151
OTU 65	sun70
OTU 66	sun76, sun82
OTU 67	sun77, sun83, sun85, sun92, sun99, ms191
OTU 68	sun95
OTU 69	ms139, ms166
OTU 70	ms141, ms167
OTU 71	ms147
OTU 72	ms170
OTU 73	ms190, ms60
OTU 74	or101, or102, or108, or50, or52, or57, or60, or68, or78, or85, or87, or91, or92, or93, or94, or95, w8, or111, or112, or113, or49, or53, or62, or69, or70, or76, or83, or107, or73
OTU 75	or103, or104, or110, or115, or51, or56
OTU 76	or114, or82, or84
OTU 77	or48, or74, or96, or80
OTU 78	or54
OTU 79	or55, or90
OTU 80	or61, or75, or77, or89, w23
OTU 81	or66, m35, m80, w40, w57
OTU 82	m110
OTU 83	m113, m116, m122, m19, m26, m27, m28, m29, m45, m50, m51, m52, m59, m61, m70, m73, m78, m8, m83, m86, m92, m94, m49
OTU 84	m114, m84
OTU 85	m22
OTU 86	m23
OTU 87	w1, w13, w15, w22, w52, w53, w62, w64, w45
OTU 88	w10
OTU 89	w38

Table A.8. Bray-Curtis community distance matrix created using relative species abundance of each OTU at each site.

site	1	2	3	4	5	6
1	0					
2	0.9666667	0				
3	0.9888889	0.9789474	0			
4	0.8799517	0.760364	1	0		
5	1	1	0.8536585	1	0	
6	1	1	0.4631579	1	1	0
7	0.8122807	0.9789474	1	0.8937071	0.9157895	1
8	0.9888889	0.9764706	0.955418	1	0.9642755	1
9	0.9444444	0.7281654	0.9894737	0.768599	0.8777778	1
10	0.9777778	0.8155447	0.9789474	0.8736842	0.9878049	1
11	0.8174129	0.8208955	0.9446976	0.8411746	0.9552239	0.9552239
12	1	1	0.754717	1	0.950069	0.7924528
13	0.9546547	0.6784342	0.9789474	0.8690103	1	1

7	8	9	10	11	12	13
0						
0.9894737	0					
0.7982456	0.9431373	0				
1	0.8526316	0.9339181	0			
0.9236449	0.9764706	0.7205638	0.9552239	0		
1	0.8397336	1	0.8635551	0.9253731	0	
0.9894737	0.9764706	0.863964	0.8736842	0.8563937	1	0

Table A.9. Euclidean distance matrix of range transformed environmental variables.

site	1	2	3	4	5	6
1	0					
2	0.8695949	0				
3	1.702244	0.8998939	0			
4	0.7286441	0.7079314	1.3301569	0		
5	0.6973383	0.7461232	1.36359	0.8570091	0	
6	1.7100052	1.0318283	0.8165846	1.2089529	1.3479628	0
7	0.4276547	1.0153468	1.8270548	0.8818884	0.5980776	1.7269333
8	1.2929926	0.8187573	0.9911456	1.1691102	0.6675918	1.0142081
9	0.7362665	0.8611946	1.6109292	1.1706616	0.7944826	1.816604
10	1.4249442	1.1370895	1.3087373	1.4080936	0.8258327	1.4654204
11	0.8413773	0.3008419	0.985407	0.6432519	0.5837984	0.9102259
12	0.9439747	0.608261	1.0013502	0.8071362	0.6996478	1.2994941
13	1.1011844	0.7177267	1.0076307	0.8939925	0.5396071	0.8584858

7	8	9	10	11	12	13
0						
1.1909713	0					
0.7739277	1.1360583	0				
1.2786844	0.509488	1.1349982	0			
0.8995517	0.6898572	0.9549881	1.0707202	0		
1.048667	0.7744424	0.8585867	0.8260396	0.6911686	0	
1.0505267	0.4345822	1.2021915	0.8468948	0.4818846	0.7826728	0

Figure A.7. Neighbor-joining tree of *rpoB* for 128 Caldwell Field isolates selected for maximal diversity. The tree is rooted with the *rpoB* sequence from *Mycobacterium smegmatis*.

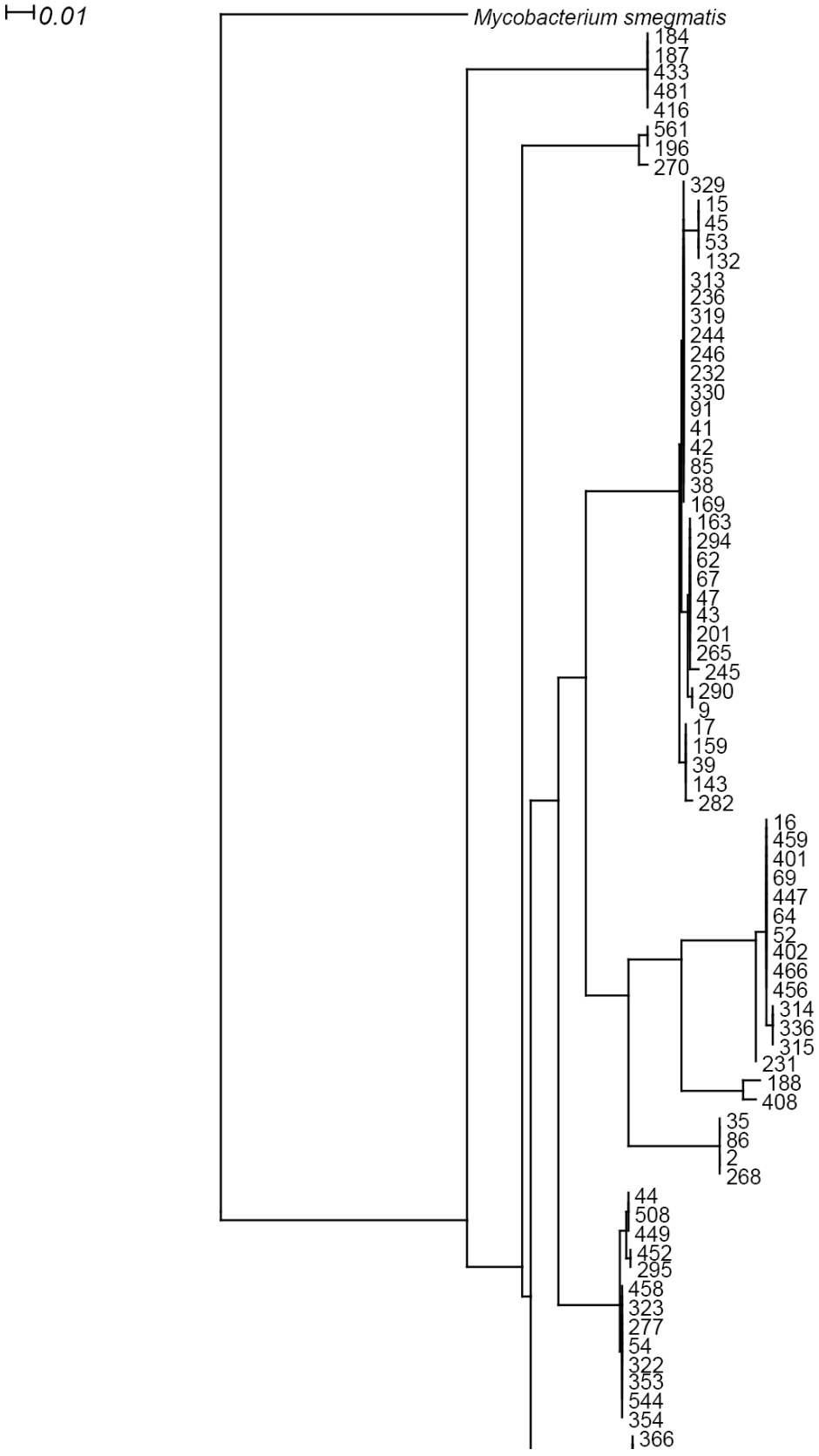

0.01
Mycobacterium smegmatis
184
187
433
481
416
561
196
270
329
15
45
53
132
313
236
319
244
246
232
330
91
41
42
85
38
169
163
294
62
67
47
43
201
265
245
290
9
17
159
39
143
282
16
459
401
69
447
64
52
402
466
456
314
336
315
231
188
408
35
86
2
268
44
508
449
452
295
458
323
277
54
322
353
544
354
366

Figure A.7 (continued)

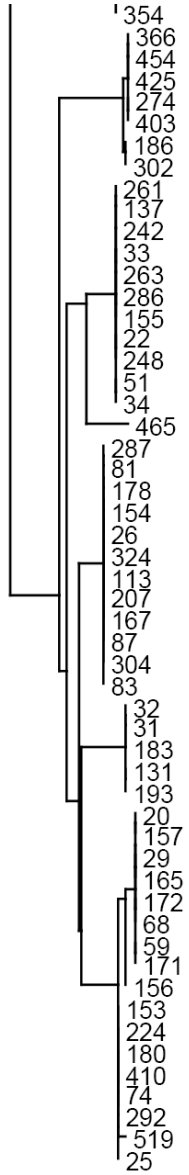


Figure A.8. Neighbor-joining tree of *trpB* for 128 Caldwell Field isolates selected for maximal diversity. The tree is rooted with the *trpB* sequence from *Mycobacterium smegmatis*.

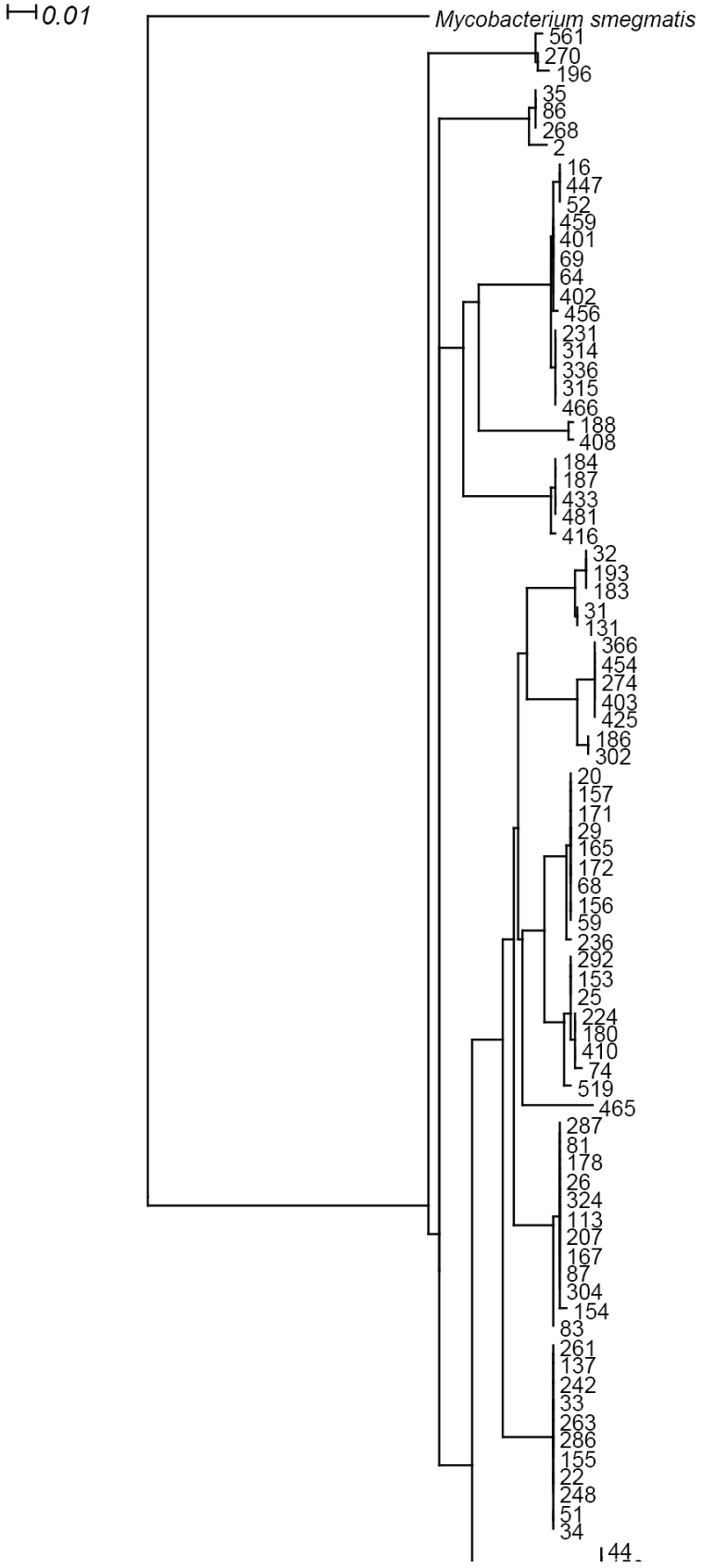

0.01
Mycobacterium smegmatis
561
270
196
35
86
268
2
16
447
52
459
401
69
64
402
456
231
314
336
315
466
188
408
184
187
433
481
416
32
193
183
31
131
366
454
274
403
425
186
302
20
157
171
29
165
172
68
156
59
236
292
153
25
224
180
410
74
519
465
287
81
178
26
324
113
207
167
87
304
154
83
261
137
242
33
263
286
155
22
248
51
34

Figure A.8 (continued)